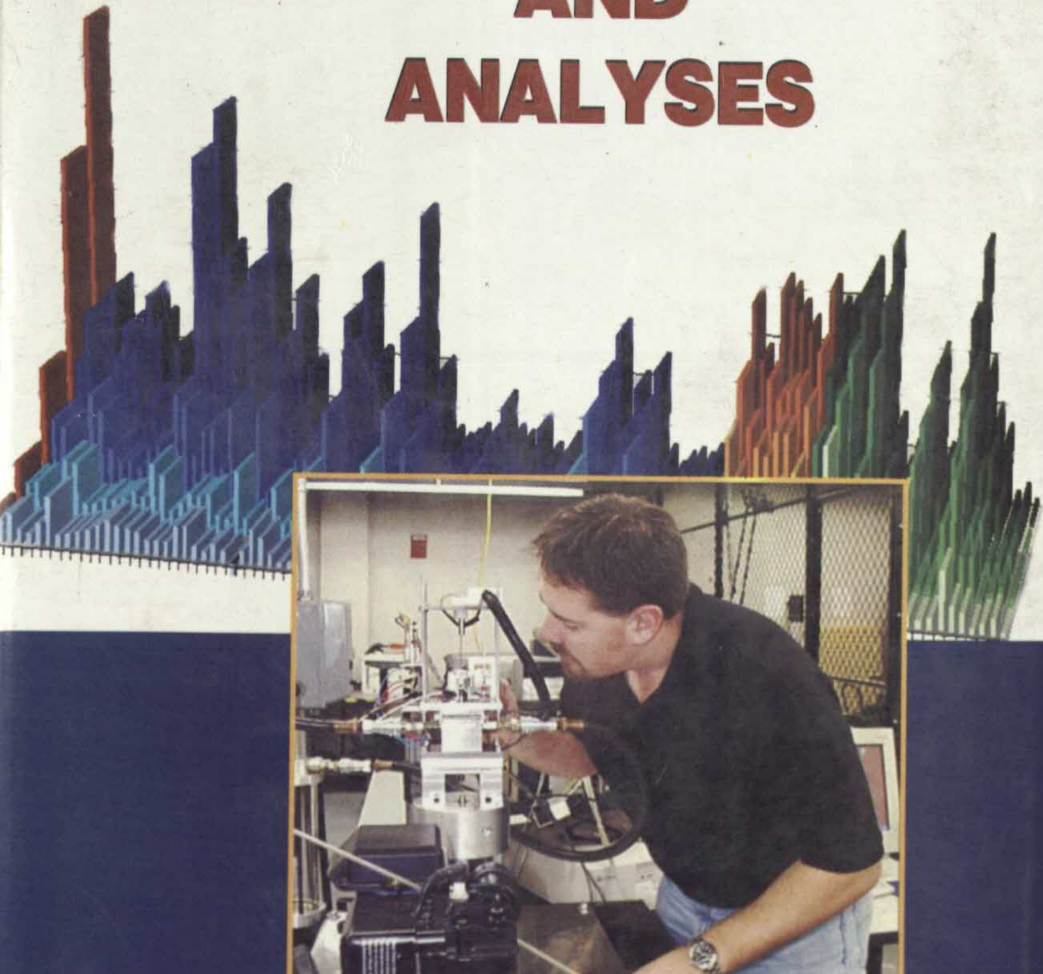
RESEARCH METHODOLOGY AND ANALYSES



R. P. SARMA
RABI N. MISRA

In the modern era of Globalisation and Computerisation, research activities in all respects has created a distinguished place not only in academic field but also in the field of Trade, Commerce, Industry and what not. Due to the development of sophisticated techniques, ideas, innovation, the idea of new thoughts has immense value after 2000. Due to scarce resources now all are interested to use their limited resources for greater productivity. To acquire this technical knowledge research is highly essential. To increase knowledge, appropriate application of ideas, resources, thoughts, techniques are required. For this research is essential. The most developed countries like USA, UK, Japan and others spend more and more on research activities.

This book has its own value to guide the innocent scholars in the right direction. This book is not only helpful for the teachers, but also for Government, the planners, industrialists, bankers, students, Non-government organisations, and others in preparing plans, budget and co-ordination of various activities of their own organisations.

CONTENTS

1. Analyses in Research
2. Time Series Analysis
3. Regression Analysis
4. Analysis of Nominal Data
5. Analysis of Ordinal Data
6. Factor Analysis
7. Discriminant Analysis
8. Cluster Analysis
9. Systems Analysis
10. Network Analysis
11. Activity Analysis
12. Dimensional Analysis
13. Meta Analysis

Rs. 750

Research Methodology
and
Analyses

Research Methodology and Analyses

By

Prof. R.P. Sarma

Prof. (Dr.) Rabi N. Misra

Library
Calcutta
23/3, B.C. Road

DPH

DISCOVERY PUBLISHING HOUSE
NEW DELHI

First Published–2006

ISBN: 81-8356-110-1

007
SAR

© Author

Published by:
DISCOVERY PUBLISHING HOUSE
4831/24, Prahlad Street, Ansari Road, Darya Ganj
New Delhi–110 002 (India)
Phone: 23279245, • Fax: 91-11-23253475
e-mail: dphtemp@indiatimes.com

[illegible]
Date.......12.3.2007
Acc. No. 12414

Printed at:
R.S. Offset Printers & Binders
New Seelampur
Delhi–110 053

Preface

In the modern era of Globalisation and Computerisation, research activities in all respects has created a distinguished place not only in academic field but also in the field of Trade, Commerce, Industry and what not. Due to the development of sophisticated techniques, ideas, innovation, the idea of new thoughts has immense value after 2000. Due to scarce resources now all are interested to use their limited resources for greater productivity. To acquire this technical knowledge research is highly essential. To increase knowledge, appropriate application of ideas, resources, thoughts, techniques are required. The most developed countries like USA, UK, Japan and others spend more and more on research activities.

This book has its own value to guide the innocent scholars in the right direction. This book is not only helpful for the teachers, but also for Government, the planners, industrialists, bankers, students, Non-government organisations, and others in preparing plans, budget and co-ordination of various activities of their own organistions.

Prof. R.P. Sarma
Prof. (Dr.) Rabi. N. Misra

Acknowledgement

In writing this book we have taken help from all corners, but we extend our thanks to Dr. Bhagabat Patro, Reader in Economics, Berhampur University and to Dr. N.C. Sahu, Head, Department of Economics, Berhampur University for their kind help and co-operation in writing this book. We are also very much thankful to Dr. Umakanta Misra, Reader in Mathematics, Berhampur University for his kind mathematical suggestions.

We are also very much thankful and grateful to Sri Tilak Wasan, Proprietor, Discovery Publishing House, New Delhi who extended his full support for publishing this book in spite of his busy schedule. We are also thankful to all the staff members of Discovery Publishing House for their kind co-operation and encouragement in Publishing this book in time.

Prof. R.P. Sarma
Prof. (Dr.) Rabi. N. Misra

Contents

Preface v
Acknowledgement vii
1. Analyses in Research 1
2. Time Series Analysis 10
3. Regression Analysis 37
4. Analysis of Nominal Data 70
5. Analysis of Ordinal Data 85
6. Factor Analysis 95
7. Discriminant Analysis 122
8. Cluster Analysis 136
9. Systems Analysis 161
10. Network Analysis 188
11. Activity Analysis 217
12. Dimensional Analysis 235
13. Meta Analysis 258
Bibliography 285
Index 289

1

Analyses in Research

Scientific and systematic search for knowledge is called research. Men's inquisitiveness is the main spring of knowledge and advancement of civilization takes place through the gradual exploration of the unknown. In other words research is an original contribution to the existing stock of knowledge providing the mankind for the new advancement of living. Even though each research study has specific purpose but the overall objective of research is to find out the truth, the truth which is hidden and not yet known.

According to the objectives of the study the researches can be classified into several types. The major types of research are as follows:

Descriptive Research: Surveys and fact finding studies conducted to know the state of affairs as it exists, are called descriptive research. In economics and social sciences this type of research is often called 'Export facto' research. In this research the researcher has no control over the variables; he can report only objectively, what had happened tries to find out the causes of the variables and their behaviour.

Analytical Research: Analytical research is slightly different from the descriptive research. In the former, the information is collected and the relationships of the variables are ascertained, but in case of the latter, in addition to it the materials so collected are critically analysed with the information already available in the field. It gives a critical conclusion with the new information got from the data obtained in comparison to the existing notions and information.

Applied Research: When a research is carried on with an objective of findings out a solution for the problem faced by a social scientist, industrialist or a business organisation it is called applied research or action research. Most of the researches carried on by the industrial houses to know the market behaviour of the consumers to evolve a marketing policy for their products, or to find out a new process of production to curtail costs are coming under applied research. The Research and Development units of the industrial organisations generally carry out these type of research.

Fundamental Research: Fundamental research is called pure or basic research. It is mainly concerned with the generalizations and with the formulation of theories. Result of this type of research is not to provide any solution for a particular problem immediately, but can be utilized for several purposes and can be developed further for solutions in several fields. In this type of research, the research is done for "gathering knowledge for the sake of knowledge". Even though the outcome of this research may not solve any immediate problem, but this would increase the knowledge of human beings and can be utilized in many field of activities in the future. In the social sciences, research studies conducted to know the human behaviour in general can be taken as the fundamental research.

Laboratory Research: Research carried out in a laboratory through experiments and by controlling the variables is called laboratory research. In social sciences experiments can also be taken in the field which are called field-setting research. A scheme is introduced in an area and its results are noted down for evaluation to know how far the experiment has been successful or not. Before implementing a scheme for the alleviation of poverty in a country, first it can be experimented in a block for its impact on the economic condition of the people. In the experiment some variables are manipulated to observe their effect on other variables.

Research Methods

Various techniques used in research are called research methods. In the research process different techniques are used at the three different levels:

1. Techniques used in the collection of data.
2. Statistical techniques used for establishing relationship among the variables.
3. Techniques used to evaluate the accuracy of the results.

Research method is generally confused with the term "Research Methodology" which has broader meaning. Research method is a technique while methodology is a science of studying how research is conducted by a researcher. It is a blue print of the research operations to be done by the researcher. The methodology of a research defines research problems, technique of sampling and data collection and technique of analyzing data. Hence research methodology involves the entire process of research while the research method is an exclusive technique adopted by a researcher at the different levels of research process.

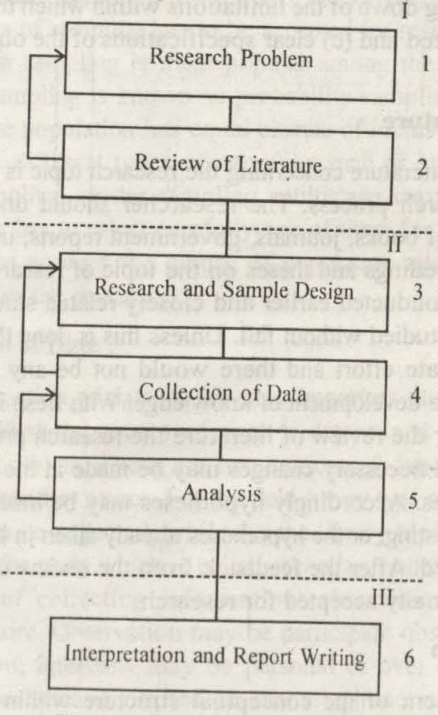


Fig. 1.1. The Research Process.

Research process can be divided into six phases and the entire process is termed as research methodology. The process is shown in the form of a flow chart in Fig. 1.1. Again the research process can be divided broadly into three parts. The first part constitutes two preliminary steps: (1) Definition of the research problem and (2) Review of literature of the research topic.

Definition of the Problem

Selection of the research problem is the first task of a researcher. Before finalizing the topic he should consult others and also experts on the subject. One normally should not select a subject which is over-done or controversial in nature and at the same time it should not be too narrow or too vague a problem. When the problem is defined in clear terms it is said the research is half done. Defining a problem involves three specific factors: (a) clear statement of the problem (b) laying down of the limitations within which the research has to be conducted and (c) clear specifications of the objectives of research.

Review of Literature

Review of literature concerning the research topic is the second step in the research process. The researcher should undertake an extensive study of books, journals, government reports, unpublished conference proceedings and theses on the topic of research subject. Similar studies conducted earlier and closely related studies on the subject must be studied without fail. Unless this is done the research may be a duplicate effort and there would not be any significant contribution to the development of knowledge. With fresh information on the topic after the review of literature the research problem may be redefined, and necessary changes may be made in the limitations and the objectives. Accordingly hypotheses may be inserted afresh if necessary for testing, or the hypotheses already taken in the problem may be eliminated. After the feedback from the review of literature the problem is finally accepted for research.

Research Design

The statement of the conceptual structure within which the research would be conducted is known as "Research design". It is the blue print for the research project which specifies the nature of collection of data and its analysis. Decisions with regards to 'what', 'which', 'where' and 'when' of the research process are clearly stated in the research design. Once the design is prepared the researcher proceeds for the execution of the research project taking guidance from the design in every consecutive steps of the research process. The design also includes the time frame and cost budgets within which the researcher has to carry on his research.

Research design includes also a sample design. As survey of an entire population is time consuming, it is meaningless to include all the similar units having same characteristics, hence only a part of the population is taken for study which is called "Sample". A sample design is a plan for obtaining a sample for a given population. The researcher has to mention here the size and the technique of sampling. After determining the size of the sample, one of the standard techniques of sampling has to be accepted for research. The method of random sampling is most popular among the researchers. This type of sampling is known as probability sampling, in which each item in the population has equal chance of inclusion in the sample. There are different types of sampling such as stratified sampling, quota sampling, cluster sampling, multistage sampling etc., and the researcher has to select carefully one of them for collection of data taking into account the nature of population and the information required for the research.

Collection of Data

Collection of data is the next important step in the research process. First the source of data is determined according to the research problem. There are two sources of data: primary sources and the secondary source. For the field survey type of research primary data has to be collected, while for the library topics published data are taken for study which is called secondary source. There are three methods of collecting information: observation, interview and questionnaire. Observation may be participant observation or group observation; interview may be personal or over the telephone, of course the latter is limited to the telephone users only. In the questionnaire method the questions may be mailed to the sample persons to be filled up by themselves, or the researcher or this nominee collects information through a schedule.

Analysis

The fifth step in the research process is analysis of data collected. Before analysis the raw data has to be processed. First the data collected are edited for correct and accurate information and cross checked to eliminate errors in collection. Sometimes coding of data is required before tabulation for speedy analysis. Finally the data are tabulated according to the nature of data and requirement of the researcher and kept ready for analysis.

Literally analysis means defining the parts of a theme and establishing their relationship in detail. In social sciences or in any other science analysis means only quantitative analysis. Since quantitative analysis has to be performed through statistical tools based on mathematical or logical properties, analyses are generally regarded as statistical analysis. Hence one can define analysis as statistical analysis or performing statistical analysis of a theme or problem. Of course the qualitative factors are also can be taken for analysis but for the purposes of statistics the qualitative data have to be converted into quantities over a given scale determined by the researcher. Sometimes the scales may be arbitrary or standardized but the quantitative statistical tools used to analysis it must be based on statistical norms.

Is economic analysis different from the analysis described in general? It seems there is no apparent difference. Statistical analysis performed for economic interpretation can be described as economic analysis. Similarly chemical analysis or political analysis can be visualized as statistical analysis performed for the purpose of analyzing chemicals or political behaviour of a country.

Adam Smith (1723-1790), the father of modern economics established economics as a separate subject of study with the publication of "Wealth of Nations" (1776); and termed it as "Political Economy". This term remained current throughout the nineteenth century until the publication of "Principles of Economics" in 1890 by Alfred Marshall. He gave a new nomenclature to economic science as "Economics" which is now accepted as the standard term. J.A.

Schumpeter (1883-1951) was not satisfied with this term, instead he preferred to call economics as "Economic Analysis". He made a good distinction between "Political Economy" and "Economics". Political economy is defined as public economic policy that exist at a given time and place float in the mind of the public. On the other hand "Economics" or "Economic Thought" is identified with ideology or general attitude or spirit through which economic problems are approached. Accordingly his monumental work on history of economic thought, published posthumously by his wife in 1954, he titled it as "History of Economic Analysis"

Professor K.E. Boulding accepted Schumpter's term "Economic Analysis" for "Economic". He defines economic analysis at the objective analysis of economic factors, since the greater part of economics "is concerned with investigating the nature of economic quantities, the relationship existing between them and the forms which determine them". Here the economic quantities mean all such quantities interest to the economists: price of commodities, production, marketing and consumption of goods and services, export and import of goods, quantity of money issued by the Central Bank etc. Collection of these quantities is the task of economic statistics and their interpretation is the principal task of economics hence it is called Economic Analysis.

In economics two types of tools are used for analysis; one set of tools exclusive to economic science and the other common to all sciences in general. The latter set of tools are described as statistical tools of analysis.

The first set of tools can be stated as economic tools. Two of the economic tools used for the purpose of theories are marginal analysis and the equilibrium analysis:

1. ***Marginalism:*** Marginal analysis of consumer behaviour is a well-established economic tool that has been used from the second half of the nineteenth century by the economists. Even though in recent years its use has been restricted but still it is not abandoned.

2. ***Equilibrium:*** Equilibrium analysis developed by the classical school was extensively used by Karl Marx for macro-level

analysis; it is still an important tool for analyzing economic problems. To-day it is further developed in the form of multiple equations to build up economic and econometric models.

The purpose of this book is to introduce few of the sophisticated statistical tools for the benefit of the researchers in economics and social sciences which are so far extensively used by the physical and biological scientists. This only familiarizes the different methods of analysis to rouse interest among the social scientists towards the modern tools. Hence the analyses explained here are not exhaustive, the researchers want to utilize these tools in their study have to learn further details about these.

The book begins with the explanation of two traditional statistical tools: the time series analysis and the regression analysis. These are given to keep the researcher at ease in the initial chapters and to create interest in the subsequent ones. The modern analyses introduced are: (1) Analysis of Nominal and Cardinal Data (2) Factor Analysis (3) Discriminant Analysis (4) Cluster Analysis (5) Systems Analysis (6) Network Analysis (7) Dimensional Analysis and (8) Meta-Analysis. Software for these analyses are now available in most of the computer manufacturers and a researcher once has a good knowledge of these analyses can utilize them by processing the data in a computer.

Collection of data and its analyses are the two most important activities in the research process which involves major part of a researcher's time, energy and finance. The final result of the research depends mainly on analysis. But no part of the research design can be said as mutually exclusive; any difficulty faced at the collection stage may fed back to the sample design or any lacuna found at the analysis level may require a revision in the type and nature of data collected in the interest of good research and better research results.

Interpretation and Report Writing

Finally the researcher has to prepare a report for what he has found from the research. The main text of the report is prepared in three parts: introduction, the analysis chapters and the conclusion.

The entire process of research from the step one to five, as shown in the block diagram, are indicated in detail in the introduction chapter. Then the main body of the report describes the different analysis of the research problem arranged in suitable number of chapters in a logical sequence. The last chapter of the report is devoted for the results of research study. In this chapter the hypotheses are tested and the researcher has to arrive at generalization to build up a theory whatever possible. The findings of the research has to be put in summary form clearly and precisely. Sometimes recommendations are prescribed for the rectification of the findings if the researcher is interested. But in applied research the recommendations form a part of the concluding chapter.

2

Time Series Analysis

Time series is one of the important techniques used by the social scientists. Economists have been the pioneers in developing the time series. It is now widely used by the sociologists, psychologists, historians and the political scientists for analyzing change over time.

Researchers come across two types of analysis of the variables in collection of data:

1. ***Cross-sectional Data:*** In this case the researchers observe a set of variables at a given point of time across space or other units of analysis. In cross-sectional data the time element is not taken into account.
2. ***Time Series Data:*** In this data variables of the same unit of analysis are observed but over a series of time points, of months, years or days. The time series analysis takes into account the change over time. Events in time move in one direction—forward; hence in this analysis data also move in one direction when examined over time. In other words the main difference between the cross section data and the time series data is that in the latter case the data are processed in the order of time period, while in the former it is not.

Two of the most important objectives of the time series analysis are:

1. The analysis shows the trend of the movement of variables over time, as in case of regression analysis, in linear form.
2. The time series analysis makes forecasts into the future from the data in the past. This is one of the important uses in

economics and management science. A time series regression analysis both explains the past and predicts the future behaviour of the variables. It informs about the peculiar mechanism which describes the evolution through time and allows the researcher to put this mechanism in forecasting the future.

In fact the time series analysis is simply a regression analysis between two sets of variables of which one is a time variable. In demand analysis the relationship of demand of goods and its price are taken as two sets of variables over a single point of time. But if demand is analysed at different points of time it becomes a time series analysis.

Finding of Linear Trend

Algebraically the general equation for a straight line in a two dimensional X-Y plane is

$$Y = a + bX$$

where Y is an endogenous variable or often referred as dependent variable: and X is the exogenous variable which often called explanatory or independent variable. The constant 'a' is the Y—intercept, that means when the value of X is zero it indicates the value of Y. The constant coefficient 'b' is the slop of the straight line, as only when slope of a line remain constant it becomes a straight line.

Making an exact equation of the above:

$$Y = 2 + 4X$$

This equation indicates that the Y intercept is 2 and slope of the straight line is positive and 4; that means when X increases by 1 unit Y increases by 4 units.

Observe Table 2.1 which presents time series data of production of cement in India. This can be plotted in a X-Y plane taking time in years as X and production of cement in tones as Y.

To know the trend over time a straight line can be fitted to know the relationship between the time and production of cement in India. This straight line can be expressed in the form of an equation, once the constant coefficients 'a' and 'b' of the trend line are known.

Table 2.1: Production of Cement

(Figures in lakh tones)

Year (X)	*Qantity (Y)*
1980-81	27
1985-86	56
1990-91	80
1995-96	108
2000-01	132
2001-02	192
2002-03	186

The most important method of finding the trend line is through the method of least squares. But roughly a quick trend line can be found for any time series data through simple methods. These methods are (1) the method of freehand, (2) method of semi-averages and (3) method of moving average, which would be discussed first before the method of least squares.

Free-hand Method

In all these simple methods first the time, whether in years months or days, is restated in terms of single numbers in ascending order shown as X before finding the trend line equation. Col. 1 of Table 2.1 is shown as from 0 to 6 in Table 2.2 below.

Table 2.2: Conversion of Time Period into X

Year	*X*	*Y*
1980-81	0	27
1985-86	1	56
1990-91	2	80
1995-96	3	108
2000-01	4	132
2001-02	5	192
2002-03	6	186

Procedure of Finding the Trend Line: Plot two points, the first and the last one or any two, and connect the two points.

The straight line joining the two points is the freehand trend line. In Table 2.2 the year 1980-81 is taken as origin and hence zero, and accordingly the subsequent years are denoted as 1, 2, 3... respectively. But any year can be taken as origin. If 1990-91 is taken as origin i.e. zero, the years previous to it can be termed as –1, and –2 respectively. The coordinates of the points and the freehand trend line is shown in Fig. 2.1. The trend line is drawn connecting the first and the last points.

According to the figure 1980-81 = 0 and 2002-03 = 6; so the co-ordinates of the two points selected are (0, 27) and (6, 186). Substituting these values into a straight line equation the freehand trend line equation can be found.

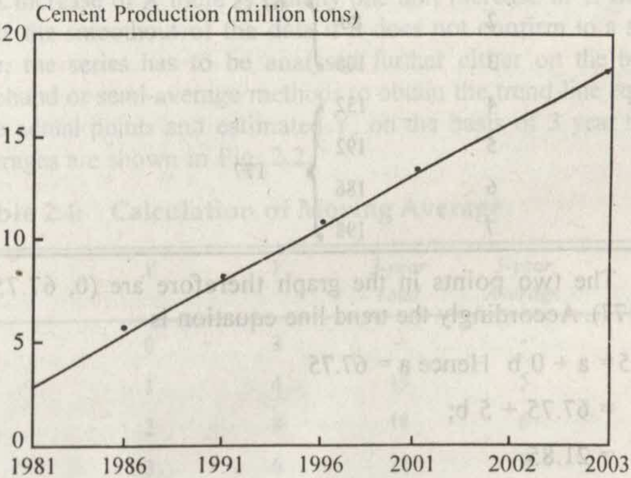


Fig. 2.1. Free Hand Trend Line

$27 = a + 0\ b$ hence $a = 27$

$186 = a + 6\ b$

$186 = 27 + 6\ b$

$186 - 27 = 6\ b$

$159 = 6\ b$ hence $b = 26.5$

Accordingly the estimated Y_c which represents the trend line is found as

$$Y_c = 27 + 26.5\ X$$

Method of Semi-average

In the method of semi-averages the series is divided into two parts, and for each part the arithmetic mean is calculated. On the basis of these arithmetic mean values of Y, the trend line is fitted. This is shown in Table 2.3.

Table 2.3: Semi Averages Table

X	*Y*	*Mean*
0	27	67.75
1	56	
2	80	
3	108	
4	132	177
5	192	
6	186	
7	198	

The two points in the graph therefore are (0, 67.75) and (5, 177). Accordingly the trend line equation is

$67.75 = a + 0\ b$ Hence $a = 67.75$

$177 = 67.75 + 5\ b;$

$b = 21.85$

Therefore $Y_c = 67.75 + 21.85\ X$

If the series has an odd number of years/days/months, as in case of Table 2.1, and cannot be divided into two equal parts, the middle value may be left out or mean may be calculated with one odd and the other as even point. If it is found that in some years the values are extreme values due to some other causes, the values of those years may be eliminated.

Method of Moving Average

In the moving average method generally there is no need of finding the trend line equation. The data is smoothed into a straight line on the basis of moving averages. The nature of moving averages depends upon the observed fluctuations in the time series data. If it is found that in the data the fluctuations are there in every three years a three-year moving average can be taken for smoothing the data. Accordingly 4, 5 or any number of years can be taken for calculating moving averages depending on the length of the series.

A 3-year moving average trend is shown in Table 2.4. Column 3 in the table gives the three year total value and the Col. 4 the averages, by dividing each value of Col. 3 by 3. The average figures are put in the middle of the three years, hence in the first and the last year there are no figures. The three year moving average in the table is representing a perfect straight line, since for everyone unit increase of X there is exactly one unit increase in Y. But even after the smoothout of the data if it does not confirm to a straight line, the series has to be analysed further either on the basis of freehand or semi-average methods to obtain the trend line equation. The actual points and estimated Y_c on the basis of 3 year moving averages are shown in Fig. 2.2.

Table 2.4: Calculation of Moving Average

X	*Y*	*3-year Total*	*3-year Average*
0	3	-	-
1	4	15	5
2	8	18	6
3	6	21	7
4	7	24	8
5	11	27	9
6	9	30	10
7	10	33	11
8	14	36	12
9	12	-	-

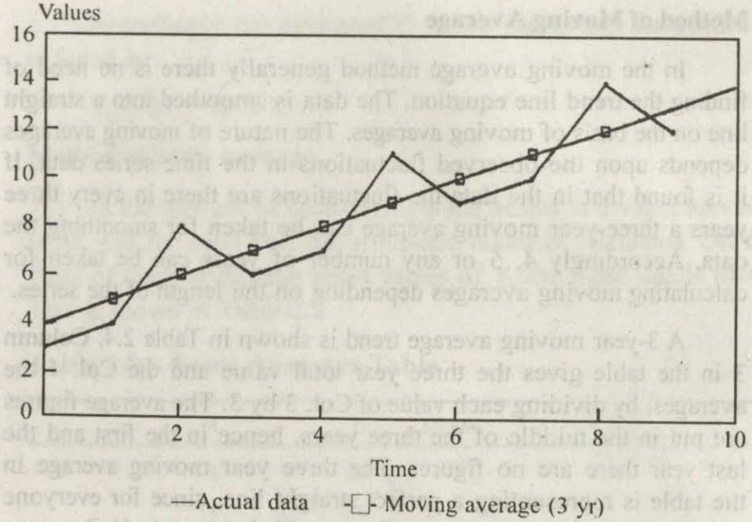


Fig. 2.2. Trend Line of Moving Averages

Time Series Regression Model

All these methods explained above are not confirming to any mathematical or statistical analysis. Hence the time series regression model is adopted for arriving at the linear relationship. The simple linear equation of two variables over time Y_t and X_t is called time series regression model.

$$Y_t = a + bX_t + e_t$$

The subscripts t indicates time, and e_t is the random disturbance term. If the sample data is perfect there is no need of this error term. Hence for simple explanation of the regression e_t is omitted.

The straight line equation is drawn on the principle that the square of the deviations of the observed points from the estimated trend line are at minimum.

The basic assumptions of this model are:

1. There is linear relationship between Y and X.
2. There is no correlation between e_t and X_t, i.e. $E(e_t X_t) = 0$
3. Mean of the error term is zero; $E(e_t) = 0$
4. Constant variance; $E(e_t^2) = d_e^2$
5. Non-autoregression; $E(e_t e_{t-m}) = 0$

This means there is no influence of previous time on the next time.

6. The error term is normally distributed.

With these assumptions the regression parameters ‘a’ and ‘b’ can be found.

Two normal equations can be derived by multiplying the linear equation with Σ and ΣX.

$\Sigma\ (Y = a + b\ X) = \Sigma Y = na + b\Sigma X$

$\Sigma X\ (Y = a + b\ X) = \Sigma XY = a\Sigma X + b\Sigma X^2$

Calculation of the parameters ‘a’ ‘b’ is easier if the sum of X that is, the time variable is denoted in such a way that its total becomes zero. Of course in regular regression analysis this is not possible because the independent variables have specific values.

Assuming that the sum of time variables is zero the two above normal variables can be reduced as:

$(\Sigma Y = na + b\Sigma X) = \Sigma Y = na$

$(\Sigma XY = a\Sigma X + b\Sigma X^2) = \Sigma XY = b\Sigma X^2$

Calculation of time series regression equation is explained with the help of a small time series presented in Table 2.5.

The total row of the Table 2.5 shows these values:

$\Sigma X = 0; \Sigma Y = 60; \ \Sigma XY = 37;$ and $\Sigma X^2 = 10$

Accordingly the regression coefficient ‘a’ can be estimated as

$(\Sigma Y = na) = 60 = 5a$ as $n = 5$; $a = 60/5 = 12$

Table 2.5: Time Series Regression Model

	X	Y	XY	X^2
	−2	5	−10	4
	−1	8	−8	1
	0	12	0	0
	1	15	15	1
	2	20	40	4
Total	0	60	37	10

The coefficient 'b' is:

$(\Sigma XY = b\Sigma X^2) = 37 = b(10);$ $b = 37/10 = 3.7$

Hence the regression equation is:

$Y_c = 12 + 3.7\ X$

In a time series regression analysis when the time variables summation is taken as zero the parameters can be easily calculated with these two simple formulas:

$$a = \frac{\Sigma Y}{n} \qquad b = \frac{\Sigma XY}{\Sigma X^2}$$

Seasonal and Cyclical Movement

The economic time series is classified in four components: (i) trend T (ii) seasonal variation S (iii) business cycle C and (iv) the irregular variation I; hence the original series O can be expressed as:

$O = T \times S \times C \times I$

Except the trend value which is in the real value the rest of the components S, C and I are indices in terms of percentages. Suppose that the O = 1890 in a particular month, of which T = 2000. In terms of percentages, S = 105, C = 100, I = 90; and the meaning of these are:

S : There is 5 per cent seasonal variation.

C: The C = 100 means that there is no cyclical change, it remains same.

I: Here 90 means that there is –10 per cent irregular variation in the month.

$O = 2.0 \times 105 \times 100 \times 90$
$\quad = 1890$

The mathematical statisticians generally use the quantitative deviations instead of percentage deviations which is comparatively easier for calculation.

Quantitative Deviations

In this method of quantitative analysis of trends the S, C, and Is are not the index but the quantitative deviations. This method is more popular among the statisticians. In this procedure decomposition is done in the form of

$$O = T + C + S + I$$

All the time series data are not in a straight line trend, and even if it can be stated as the linear trend it will be far from the real figures. The next most important time series data is in the form of seasonal changes or systematic fluctuations over time. From the original data when the trend T is removed it is called stationary time series. When the linear trend is not there the data may fluctuate below and above the trend line as shown in Fig. 2.3 and if one is interested in the deviations only the trend line can be set as zero horizontal axis or any other constant horizontal line as shown in Figs. 2.3 and 2.4 respectively. In the Fig. 2.3 the deviations are shown in terms of the +ve or –ve quantities, but in the Fig. 2.4 the seasonal variations are shown as higher or lower quantitative values.

The seasonal fluctuations can be divided into two series:

1. Periodic Series or Seasonal Variation
2. Oscillatory Series or Business Cycles

Seasonal Variations: A series that repeats itself regularly in intervals of a specific period is called periodic series. The number of time intervals (days, weeks, months or years) is called the period of the series. In the Fig. 2.4 the time distance between 'a' and 'b' is called the period of the series. The range between the upper and the lower extreme values of the series which is being repeated, i.e. the difference between 'c' and 'd' is called the amplitude of the series. In the seasonal variation the period and the amplitude remains constant. A periodic series exists only theoretically because the repetitions are not always uniform in nature.

Oscillatory Series: A series that repeats itself but which does have a fixed period is called oscillatory series. Here there is neither a fixed period nor any fixed amplitude of the series.

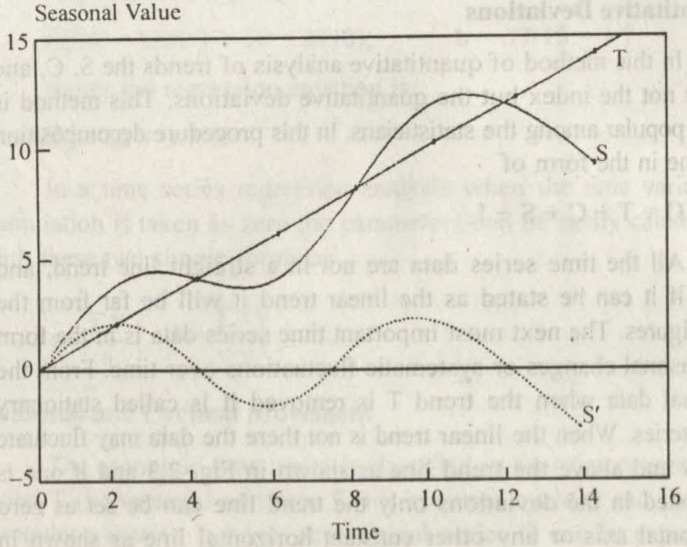


Fig. 2.3 Seasonal Variation

In both the series there are certain systematic causes for the fluctuations, but if there are no such causes it is known as random series.

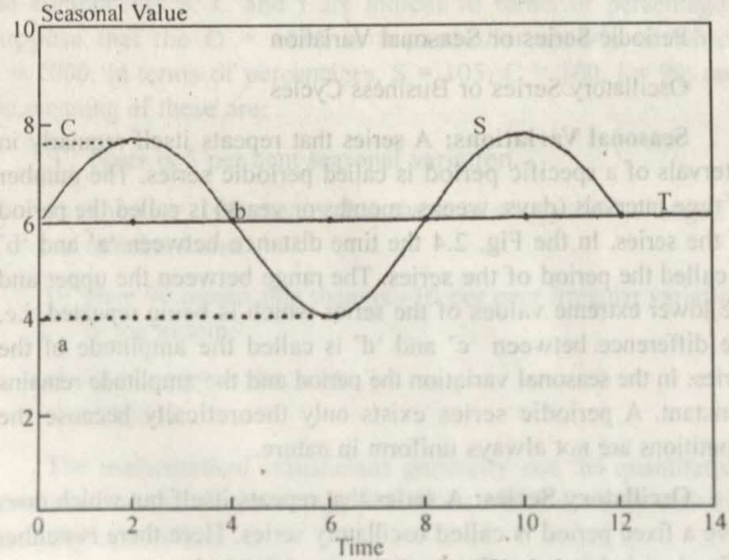


Fig. 2.4 Measurement of Seasonal Variation

Measuring of Seasonal Variations

The values of T, C, and I have to be eliminated from the series to find out the seasonal variation. For example, during a year the demand for soft drinks fluctuate from season to season; depending on the nature of the weather. In order to eliminate 'I' the averages of two years or three years for each month is calculated. Since the random disturbance of the different years are assumed to be independent of each other, the irregular variations due to the random disturbances will cancel each other. Elimination of 'I' through the method of simple averages is demonstrated with the help of Table 2.6.

Table 2.6: Method of Simple Averages

Month	*2002*	*2003*	*Total (1+2)*	*A.M.*	*T*	*A.A.M.*	*S*
	1	*2*	*3*	*4*	*5*	*6*	*7*
Jan.	560	780	1340	670	0	670	97.4
Feb.	500	720	1220	610	5	605	88.0
Mar.	450	670	1120	560	10	550	80.0
April	420	660	1080	540	15	525	76.3
May	420	630	1050	525	20	505	73.5
June	480	660	1140	570	25	545	79.3
July	590	730	1320	660	30	630	91.6
Aug.	750	860	1610	805	35	770	112.0
Sept.	860	970	1830	915	40	875	127.3
Oct.	900	980	1880	940	45	895	130.2
Nov.	900	950	1850	925	50	875	127.3
Dec.	850	870	1720	860	55	805	117.1
Total	7680	9480	—	—	—	8250	1200
Average	640	700	—	—	—	687.5	100

The averages of two years shown in column 4 eliminates the irregular fluctuations in the series. If averages of sufficiently more

Date 12.3.2007
Loan No 12419
15/3.

number of years are taken the business cycles cancel out the cyclical effects. Hence the column 4 figures eliminate both I and C, that means there remains only T and S in the series. From this column when the trend value T (Col. 5) is deducted finally one gets S or the seasonal variation. The Col. 6 is the adjusted arithmatic mean after the deduction of the T value. Based on the data from Col. 6, S is computed as percentage of each month's data to the average of the monthly data.

For example, for the month of January 1989 the S value (97.4) is computed in the following procedure.

(670/687.5) 100 = 97.45, this means 2.55 per cent less in the month of January to the monthly averages. Accordingly every month's percentages are calculated and presented as seasonal data in Col. 7 of the Table 2.6. This seasonal variation can be shown in a graph, where S fluctuates around the trend line or around horizontal axis as presented in Figs. 2.3 and 2.4 earlier.

The trend line T is calculated on the basis of annual monthly average figures and shown in Col. 5 in the table. The trend line for the monthly averages is calculated with the help of Table 2.7. Five years average monthly data are shown as Y in the table, while time is taken as X, as usual, with the year 1988 as zero.

Table 2.7: Calculation of Monthly Trend Line

Year	X	Y	XY	X^2
1996	–2	520	–1040	4
1997	–1	580	–580	1
1998	0	540	0	0
1999*	1	640	640	1
2000	2	790	1580	4
Total	2	3070	600	10

*The monthly average figures shown in this table for the years 1999 and 2000 are already shown in last row of Table 2.6, but data for the other earlier years are not shown there.

The regression coefficients are calculated with the information of Table 2.7.

$a = 3070/5 = 614$ and $b = 600/10 = 60$

Hence the regression equation is $Y = 614 + 60X$

But $b = 60$ is the annual figure whereas the monthly figure is only 5. (60/12 = 5). This means the value of every month increases by 5. If January is taken as zero, February = 10, March = 15 and so on which is shown as monthly trend in Col. 5 of Table 2.6, at T value.

Ratio to Moving Average Method

The main assumptions for this model are:

1. The seasonal variation is for a 12 month period.
2. The shape of the variation is same for each year.
3. Irregular variation I is independent for different periods.
4. The twelve month moving average is T × C.

First the SI is found out by eliminating T × C from the series according to the following formula:

$$S \times I = \frac{S \times C \times S \times I}{T \times C}$$

The elimination process is explained in Table 2.8. In the table col. (3) is the 12 month moving average. This average is calculated by adding 12 months data from July 1955 to June 1996 and dividing it by 12; the first figure in the column is then 90.7/12 = 7.57. This figure has to be put in the middle of the 12 months in the table, i.e. in between the months of December 1995 and January 1996. Then from August 1995 to July 1996, the 12 month moving average is put next between the months January and February 1996 and so on.

Since in between the two months figure is not appropriate for calculation the centring of figure is done. For centring two-month moving average is calculated for the figures shown in column (3).

In this process two of the 12 month moving average are added and average is calculated by dividing it with two and put in the centre of the two figures which comes directly against a month in the col. (4). For example, the first two figures in the column (3) are 7.57 and 7.62. Average of the two figures is 7.60; and this figure is put against the month of January 1956. The other figures in col. (4) are also calculated accordingly.

The seasonal index figures are calculated on the basis of the figures in columns (2) and (4). The col. (2) figures are expressed as percentage of col. (4) and put as seasonal index in the col. (5). In the table 30 months data are provided in the col. (2), but because of 12 month moving average and again 2 month moving average for centering only 20 month's data are finally put in the columns (4) and (5).

Removal of I

From the SI data of col. 5 of Table 2.8, the irregular fluctuations have to be removed to get finally the seasonal variations. In Table 2.9, the figures are all S × I figures of respective years. Of course all these figures are not shown in the Table 2.8 which have been calculated separately. The fifth row of Table 2.9 finally shows the seasonal index.

The total of the averages of the twelve months here becomes 1206 which should be 1200. Hence Adjusted Mean has to be calculated instead of simple mean in order to eliminate the extra figure which is here 6 as shown in the table. The figure 1206 comes to 100.5 per month which should be converted to 100. Accordingly each month's index is calculated from 100.5 in order to make it for 100. For example the Adjusted Mean for April in the table is calculated as: (105/100.5) × 100 = 104.

Other Forms of Trend Lines

Even though the linear trend is most popular among the economists and businessmen there are also other forms of trend lines. When an original time series is plotted in a graph it may show that a parabola fits more adequately than a straight line. The simplest parabola is that of a second degree one and the equation is in the form of:

Table 2.8: Elimination of Trend and Cycles

Year	*Month*	*TSCI*	*12 months moving average*	*12 months moving average centred*	*SI in per cent [(col. 2/4) × 100]*
1995	July	8.9	—	—	—
	Aug.	9.2	—	—	—
	Sept.	7.3	—	—	—
	Oct.	6.4	—	—	—
	Nov.	5.8	—	—	—
	Dec.	6.2	7.57	—	—
1996	Jan.	6.4	7.62	7.60	84
	Feb.	6.6	7.58	7.60	87
	March	7.9	7.51	7.55	105
	April	7.9	7.52	7.52	105
	May	7.8	7.53	7.53	116
	June	9.4	7.52	7.53	125
	July	9.6	7.56	7.54	127
	Aug.	8.7	7.50	7.53	116
	Sept.	5.6	7.48	7.49	86
	Oct.	6.5	7.49	7.49	87
	Nov.	5.9	7.52	7.51	79
	Dec.	6.1	7.48	7.50	81
1997	Jan.	6.9	7.50	7.49	92
	Feb.	5.9	7.48	7.49	79
	March	7.6	7.50	7.49	102
	April	8.1	7.53	7.52	108
	May	9.0	7.47	7.50	120
	June	9.0	7.46	7.47	121
	July	9.8	7.45	7.46	131
	Aug.	8.4	7.48	7.47	112
	Sept.	6.8	—	—	—
	Oct.	6.8	—	—	—
	Nov.	5.2	—	—	—
	Dec.	6.0	—	—	—

Table 2.9: Seasonal Index

Year	*Jan.*	*Feb.*	*March*	*April*	*May*	*June*	*July*	*Aug.*	*Sept.*	*Oct.*	*Nov.*	*Dec.*	*Total*
1956	84	87	105	105	116	125	127	116	86	87	80	81	1198
1957	92	79	102	108	120	121	131	112	91	91	70	81	1198
1958	92	86	100	102	118	130	131	105	97	91	78	92	1222
A.M.	89	84	102	105	118	125	130	111	91	90	76	85	1206
Adj M.	89	84	101	104	117	124	129	110	91	90	76	86	1200

$Y_c = a + bX + cX^2$

The coefficients of the equation can be found by the method of least squares on the basis of the three normal equations:

$$\Sigma Y = na + b\Sigma X + c\Sigma X^2$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 + c\Sigma X^3$$

$$\Sigma X^2Y = a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4$$

When the origin is shifted to the mean values of the series and $\Sigma X = 0$; then the normal equations are simplified to:

$$\Sigma Y = na + c\Sigma X^2$$

$$\Sigma XY = b\Sigma X^2$$

$$\Sigma X^2Y = a\Sigma X^2 + c\Sigma X^4$$

On solving these three normal equations the coefficients 'a', 'b' and 'c's can be known, to fit into the second degree equations.

Differences of the Series

The first basic question in time series analysis is how to know that the series fits into a linear, parabolic of second degree or to the higher degrees? There is one simple method to detect the nature of the trend line present in a series. It can be determined by investigating the differences of the series.

$$\Delta Y_i = Y_i - Y_{i-1}$$

This is known as first difference. In the linear trend the first difference is equal to the parameter b. This is explained in the table below:

X	$Y_c = a + bX$	$\Delta Y_i = Y_i - Y_{i-1}$	$\Delta Y_i^{\,2}$	
0	a	—		
1	a + b	(a + b) – a	= b	= 0
2	a + 2b	(a + 2b) – (a + b)	= b	= 0
3	a + 3b	(a + 3b) – (a + 2b)	= b	= 0
…	… …	……	= b	= 0

The second difference of a series is:

$$\Delta^2 Y_i = \Delta Y_i - \Delta Y_{i-1}$$
$$= b - b = 0$$

In the linear case the second differences are always zero.

The first and second differences of a second degree parabola are shown in the following table:

X	$Yc = a + bX + cX^2$	ΔY_i	$\Delta^2 Y_i$
0	a	—	—
1	a + b + c	b + c	—
2	a + 2b + 4c	b + 3c	2c
3	a + 3b + 9c	b + 5c	2c
4	a + 4b + 16c	b + 7c	2c
5	a + 5b + 25c	b + 9c	2c

The first and second differences can be shown algebraically in the following:

$$\Delta Y_i = (a + bX_i + cX^2_i) - (a + bX_{i-1} + cX^2_{i-1})$$
$$= b\,(X_i - X_{i-1}) + c(X^2_i - X^2_{i-1})$$
$$= b + c(X_i - X_{i-1}) \text{ where } (X_i - X_{i-1}) = 1$$

The second difference is:

$$\Delta^2 Y = \Delta Y_i - \Delta Y_{i-1}$$
$$= c\,(X_i - X_{i-2})$$
$$= 2c$$

Now the general principle can be the nth difference of nth degree parabolic equation is constant. On this basis whether a series can be taken up for a linear trend or a parabolic equation is determined. Hence the two simple principles for the selection of trend line are:

1. If the first differences of a series are approximately constant the linear trend line can be taken for analysis.
2. On the otherhand if the second differences are approximately constant a second degree parabolic equation can be fitted.

The Exponential Trend Line

Growth of population, national income and other phenomena which grow geometrically, fitting of the exponential trend line is more suitable for analysis.

The equation is:

$Y = ab^{\alpha}$

When the value of $b > 1$, the value of Y will increase infinitely, as X times becomes large; but when $b < 1$, it will approach zero.

The above equation if expressed in terms of logarithms it becomes a linear function.

$\log Y = \log a + X \log b$

When the series is plotted in a semi-log scale a straight line is formed in the graph. The slope of the line is log b and it is taken as the rate of increase of Y.

The two normal equations for the exponential trend line in the log form are:

$\Sigma \log Y = n \log a + (\log b)\, \Sigma X$

$\Sigma X \log Y = (\log a)\, \Sigma X + (\log b)\, \Sigma X^2$

Accepting $X = 0$, the formula for the regression coefficients are:

$$\log a = \frac{\log Y}{n}$$

$$\log b = \frac{\Sigma X \log Y}{\Sigma X^2}$$

Modified Exponential Trend

Without using the logs the exponential trend can be found by adding a constant to the exponential equation given above as under:

$Y_c = k + ab^X$

When a constant is added it becomes the uppermost or lowermost approach of the trend line, as the case may be. Once the constant is known and the regression coefficients 'a' and 'b' are estimated the trend line can be plotted.

For example when the estimated trend line is:

$Y = 32 + (-1.6)(0.5)X$

The trend line can be plotted and the value of Y can be obtained as under:

X :	−1	0	1	2	3	4
Y :	0	16	24	28	30	31

Y here approaches the constant 32 which is the value of the constant in the equation. According to the nature of the coefficients the shape of the curves can be divided into four. They are:

(a) $a < 1$ and $b < 1$: The curve increases from left to right and approaches the constant at the top.

(b) $a < 1$ and $b > 1$: The trend line moves from left to right downwards and approaches X axis from the constant.

(c) $a > 1$ and $b < 1$: The trend line moves from left to right downwards and approaches the constant at the lowest level.

(d) $a > 1$ and $b > 1$: The line moves upwards from the constant infinitely.

These four trend lines are shown in the four sub-graphs of Fig. 2.5.

The first differences of the series are explained below:

$$1 = Y_{c2} - Y_{c1} = ab^2 - ab = ab(b-1)$$

$$2 = Y_{c3} - Y_{c2} = ab^3 - ab^2 = ab^2(b-1)$$

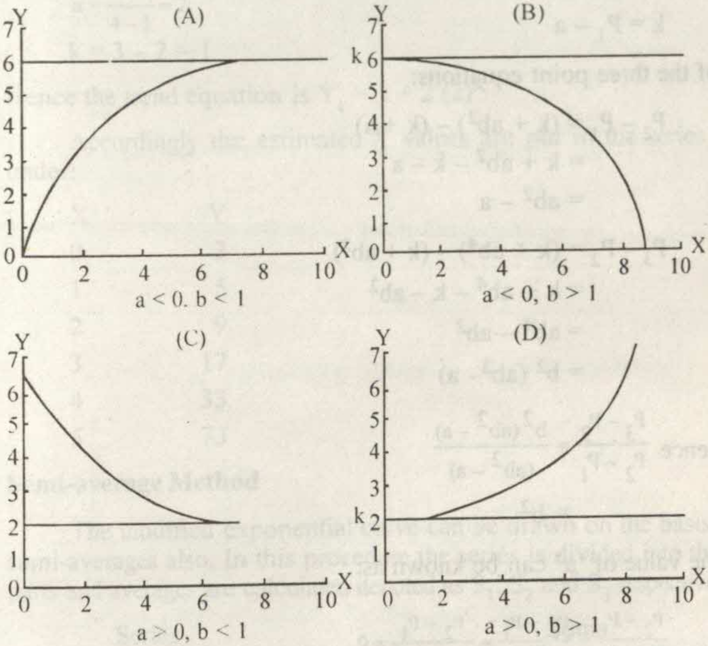
$$3 = Y_{c4} - Y_{c3} = ab^4 - ab^3 = ab^3(b-1)$$

The ratios of the first differences:

$$2/1 = \frac{ab^2(b-1)}{ab(b-1)} = b$$

$$3/2 = \frac{ab^3(b-1)}{ab^2(b-1)} = b$$

Since the trend line is not linear the method of least squares is not applicable in this case. A selective three point method can be used as a simple procedure to find out the exponential trend. In this case the ratios of the first differences are constant b which can be measured as measure of the rate of growth.



$a < 0.\ b < 1$ $a < 0,\ b > 1$

$a > 0,\ b < 1$ $a > 0,\ b > 1$

Fig. 2.5 Exponential Trend Lines

The three point method is explained below for the trend line shown in the graph 4 of Fig. 2.6; in which the parameters ‘a’ and ‘b’ are taken as a 0 and b 1.

First select 3 points in which $X = 0$, $X = 2$, and $X = 4$ which are denoted as P_0, P_2 and P_4. Accordingly the 3 points equations are:

$$P_1 = k + a$$

$$P_2 = k + ab^2$$

$$P_3 = k + ab^4$$

In the three unknowns and the three equations the value of the unknowns can be found in the following formulas:

$$b^2 = \frac{P_3 - P_2}{P_2 - P_1}$$

$$a = \frac{P_2 - P_1}{b^2 - 1}$$

$$k = P_1 - a$$

Of the three point equations:

$$\begin{aligned} P_2 - P_1 &= (k + ab^2) - (k + a) \\ &= k + ab^2 - k - a \\ &= ab^2 - a \end{aligned}$$

$$\begin{aligned} P_3 - P_2 &= (k + ab^4) - (k + ab^2) \\ &= k + ab^4 - k - ab^2 \\ &= ab^4 - ab^2 \\ &= b^2\,(ab^2 - a) \end{aligned}$$

Hence $\frac{P_3 - P_2}{P_2 - P_1} = \frac{b^2 (ab^2 - a)}{(ab^2 - a)}$

$$= b^2$$

The value of 'a' can be known as:

$$\frac{P_2 - P_1}{P_2 - P_1} = \frac{P_2 - P_1}{ab^2 - a} = \frac{P_2 - P_1}{a(b^2 - 1)} = 0$$

$$a = \frac{P_2 - P_1}{b^2 - 1}$$

The calculation of the trend equation is explained with the help of an example below:

Year	X	Y
1995	0	3...... .... P_1
1996	1	7
1997	2	9..................... P_2
1998	3	21
1999	4	33...................... P_3
2000	5	70

$$b^2 = \frac{33-9}{9-3} = \frac{24}{6} = 4; b = \sqrt{4} = 2$$

$$a = \frac{9-3}{4-1} = 2$$

$$k = 3 - 2 = 1$$

Hence the trend equation is $Y_c = 1 + 2\ (2)^X$

Accordingly the estimated Y values are put in the series as under:

X	Y
0	3
1	5
2	9
3	17
4	33
5	73

Semi-average Method

The modified exponential curve can be drawn on the basis of semi-averages also. In this procedure the series is divided into three parts and averages are calculated denoted as S_1, S_2 and S_3 respectively.

Series		Sums
$Y_0 = k + a$ $Y_1 = k + ab$	$\Big\} = S_1$	$S_1 = 2k + a\ (b + 1)$
$Y_2 = k + ab^2$ $Y_3 = k + ab^3$	$\Big\} = S_2$	$S_2 = 2k + ab^2\ (b + 1)$
$Y_4 = k + ab^4$ $Y_5 = k + ab^5$	$\Big\} = S_3$	$S_3 = 2k + ab^4\ (b + 1)$

In the above three equations the three unknowns are k, a and b. Solving for these unknowns:

$$b^2 = \frac{S_3 - S_2}{S_2 - S_1}$$

$$a = \frac{S_2 - S_1}{(b^2 - 1)(b + 1)}$$

$$2k = S_1 - \left[\frac{S_2 - S_1}{b^2 - 1}\right]$$

As per the earlier example:

$$\begin{array}{cc} X & Y \\ 0 & 3 \\ 1 & 7 \end{array}\Bigg\} \quad S_1 = 10$$

$$\begin{array}{cc} 2 & 9 \\ 3 & 21 \end{array}\Bigg\} \quad S_2 = 10$$

$$\begin{array}{cc} 4 & 32 \\ 5 & 70 \end{array}\Bigg\} \quad S_3 = 102$$

According to these values the coefficients are:

$$b^2 = \frac{102 - 30}{30 - 10} = \frac{72}{20} = 3.6$$

$$2k = 10 - \left(\frac{30 - 10}{3.6 - 1}\right) = 2.3; \text{hence } k = 1.15$$

$$a = \frac{30 - 10}{(3.6 - 1)(1.9 + 1)} = 2.65$$

The estimated Y's equation now becomes:

$Y_c = 1.5 + 2.65\ (1.9)^X$

Serial Correlation

One of the important assumptions of the time series analysis is that the observations are statistically independent. In other words the observed figures of one year is not dependent on other year's figures. When the time interval is long there may not be any dependence but in case of time interval is short they may be dependent. The nature of the measurement of dependency is known as that of independence.

Auto-correlation and serial correlation is generally used interchangeably. But when a distinction is made the term auto-correlation is made to indicate the population serial correlation coefficient and serial correlation to indicate the sample serial correlation coefficient.

Given a time series the correlation between successive terms of the series is defined as:

$$r_i = \frac{\mathrm{Cov}\,(X_i, X_{i+1})}{\sqrt{\mathrm{Var}\,(X_i)\ \mathrm{Var}\,(X_{i+1})}}$$

Computation of serial correlation is explained below with the help of Table 2.10.

Table 2.10: Computation of Serial Correlation

	X_i	X_i^2	$X_i X_{i+1}$	$(X_{i+1} - X_i)^2$
	2	4	10	9
	5	15	20	1
	4	16	24	4
	6	36	48	4
	8	64	16	36
Total	25	145	118	54

The computation formula for the serial correlation is:

$$r_i \frac{\Sigma X_i X_{i+1} - \frac{(\Sigma X)^2}{n}}{\Sigma X_{i^2} - \frac{(\Sigma X)^2}{n}}$$

On the basis of data presented in the table above the serial correlation coefficient is:

$$r_1 = \frac{118-(25^2)/5}{145-(25^2)/5} = -0.35$$

Test of Significance

How to know that the serial correlation coefficient is significant or not? R.L. Anderson developed a table giving the critical values of r at 5 and 1 per cent significant level, involving for each n value of the series. If the value of r_1 exceeds the corresponding value in the table one can conclude that serial correlation exists in the population.

In the above example the serial correlation coefficient is $r_1 = -0.35$ and the corresponding value in the Anderson's table is higher at –0.753. Hence it can be concluded that there is no serial correlation in the population.

3

Regression Analysis

One of the important objectives of the social scientists is to find out relationship between two sets of attributes. An economist wants to know the nature of relationship between income earnings of an individual with that of his educational qualification; or between investment of funds and production in a firm etc. Answers for these type of questions can be explained with the help of regression analysis.

In analysing data of two variables, such as income and educational qualifications of the individuals, income can be taken as variable Y and educational qualification as variable X. In physical sciences relation between two variables have exact relationship but in economics and other social sciences the relationships are not exact; for instance, the relationship between income and expenditure differs with regards to time, place and level of economic status of a person.

In regression analysis a researcher is interested in two aspects of relations:

1. In first case the researcher is interested to know the relationship per se, or the magnitude of relationship.
2. In second case he is interested to estimate one of the variables on the basis of relationship, given the value of the other variable. In other words if one wants to know what would be the income of an individual, given his educational qualification this can be known by a regression analysis of the two sets of the variables.

The magnitude of relationship can be known by computing correlation coefficient r for the two sets of variables. But for the second case of relationships regression analysis has to be made. In this chapter our main aim is to know the technique of regression in which one set of variables can be estimated with the help of other sets of variables. Out of the two sets of variables, X and Y, the variable Y has to be estimated, which is called dependent variable or 'effect' and the X is called independent variable or the 'cause'. In simple terms it can be said that Y depends upon X, or Y is a function of X. The functional relationship between the two variables may be linear or non-linear. When there is a linear relationship between the two it means there is a constant relationship throughout. Constant relationship means, geometrically it is a straight line, in the X-Y plane; hence it is called linear relationship between the two variables. Mathematically the functional relation between the two variables in a linear equation is:

$$Y = a + bX$$

In the equation 'a' and 'b' are the constant coefficients which are called regression coefficients. The relationship is shown in Fig. 3.1. Here the Y axis measures the Y variable and the X axis measures X variable. The straight line SS′ indicates the relationship between the two variables.

The Intercept

The constant 'a' is called intercept, because it indicates the point where the regression line 'intercepts' the Y-axis. As it can be seen in the Fig. 3.1, it measures the value of Y when X equals zero. In our example of income-education relationship it can show the average income of a person when there is no educational qualification of a person. If the curve SS′ is taken as the total cost of a firm, 'a' is taken as the minimum constant cost of a firm, to which variable costs are added to know the total cost. Hence when X is zero there cannot be any negative value of Y.

The Slope

The regression coefficient 'b' in the equation is the slope of the straight line. It indicates the average changes in Y associated with a unit change in X. This slope or the ratio of dy/dx remain constant whatever the value of X one takes. The variable X is taken

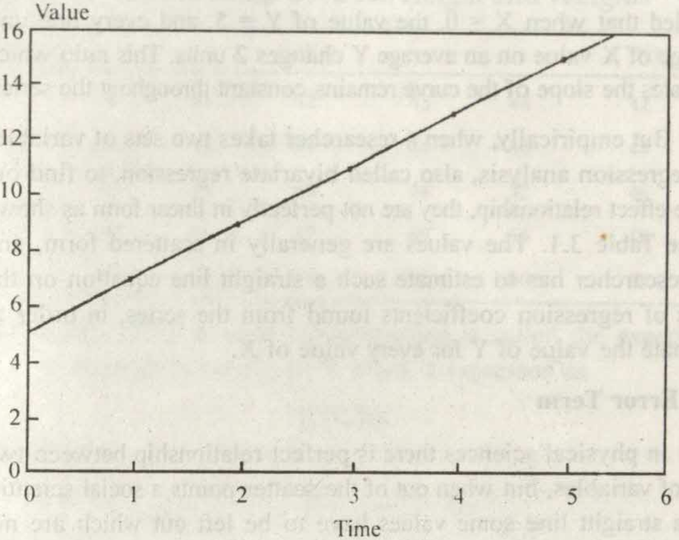


Fig. 3.1. A Perfect Linear Curve

as fixed variable or non-random variable while the Y's are the random variables. In other words the slope estimate suggests that the average change in Y caused by a unit change in X.

A simple example of linear relationship between Y and X is shown in Table 3.1.

Table 3.1: Perfect Linear Relationship

X	Y
0	5
1	7
2	9
3	11
4	13
5	15

This perfect linear relationship between the two variables can be expressed in the form of an equation of the straight line:

$$Y = 5 + 2X$$

In this equation 5 is the Y intercept and 2 is the slope. It can be verified that when $X = 0$, the value of $Y = 5$, and every one unit change of X value on an average Y changes 2 units. This ratio which indicates the slope of the curve remains constant throughout the series.

But empirically, when a researcher takes two sets of variables for regression analysis, also called bivariate regression, to find out cause-effect relationship, they are not perfectly in linear form as shown in the Table 3.1. The values are generally in scattered form, and the researcher has to estimate such a straight line equation on the basis of regression coefficients found from the series, in order to estimate the value of Y for every value of X.

The Error Term

In physical sciences there is perfect relationship between two sets of variables, but when out of the scatter points a social scientist fits a straight line some values have to be left out which are not explaining the straight line relationship, they are called the error term or the stochastic disturbance term. For example, there is perfect relationship between the temperatures of Farenheit and Celcius, taking Y and X respectively in a linear form as:

$$Y = 32 + 9.5 X$$

Here there is no question of error term because for a Celcius degree there is exactly a F degree. But relations in social sciences a particular X, since only an average of Y is taken for an actual unit of X, will not equal to a real Y. So to get the real value of Y one has to add the error term 'e' to the average Y. Since the estimated straight line only indicates the average Y, for the real Y the equation is:

$$Y = a + bX + e$$

Relationship between height X and weight Y is given in Table 3.2.

For each value of X there are 5 observations, which are called sub-population and entire observation of 25 pairs are called population.

Table 3.2: Relationship Between Height and Weights

X (inches)	*Y (Kg)*					*E (Y/X)*
50	40	41	42	43	44	42
51	41	43	44	46	46	44
52	41	44	45	48	52	46
53	43	46	47	49	55	48
54	44	46	49	51	60	50

The average of Ys in each sub-population is called the expected value of Y for given heights of X which is expressed as

$$E\ (Y/X)$$

Accordingly:

$$E\ (Y/X) = 50 = \frac{40+41+42+43+44}{5} = 42$$

and these figures are shown in the last column of the Table 3.2. But we know that one value of the sub-population is Y = 40, when X = 50. Hence the expected value of Y is less by 2 as

$$Y - E\ (Y/X = 50) = 42 - 40 = 2$$

This deviation 2 is the error term, accordingly every real term can be expressed as

$$Y = E\ (Y/X) + e$$

So Y = 40 can be expressed as

$$Y = E\ (Y/X = 50) + e$$

$$= 42 + e$$

The 'e' value can be positive or negative, here e value is negative i.e. –2, so

$$Y = 42 + (-2)$$

$$= 42 - 2 = 40$$

Hence the regression model is consisting of three following equations:

$$Y = E\ (Y/X) + e \qquad (1)$$

Since E (Y/X) falls on straight line i.e. E (Y X) = a+bx for simplicity this can be written as

$$U_{yx} = a + bX \qquad \text{................ (2)}$$

This is called the regression curve of Y on X and shows the expected value of Y on the independent value of X. This can be simply called regression function.

On the basis of the above two equations the equation (3) can be expressed as:

$$Y = a + bX + e \qquad \text{.... (3)}$$

The coefficients a and b are called regression coefficients.

Principle of Least Squares

In social sciences since the relationship is not exact, the scattered points of relationship has to be fitted into a straight line. Here the main question is what should be the basis of fitting the line? The principle is to draw a straight line through the scatter points such that the deviations from the mean value of Y with the real value of Ys should be minimum. Since the deviations may be positive or negative sometimes it may cancel each other and may not give true picture, hence the squared deviations are taken or the squared errors are taken for minimisation to fit the straight line. The line which minimises the sum of the squares of the errors, SSE is

$$SSE = \Sigma(Y \quad \hat{Y})^2$$

$\hat{Y}$ is the estimated Y for each value of X. To draw a straight line on the basis of the least squares to fit the empirical data the regression coefficients 'a' and 'b' have to be estimated. It can be shown that the sum of the squares is minimum or 'least' when the coefficients a and b are calculated as follows:

$$b = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{\Sigma(X - X)^2}$$

$$a = \overline{Y} - b\overline{X}$$

These values of a and b are the least square estimates.

When the straight line is expressed as:

$$Y = A + BX$$

A and B are called population regression coefficients. But if the equation is expressed as

$$Y = a + bX$$

the coefficients a and b are called sample regression coefficients and hence the estimated Y on the basis of methods of least squares can be written as

$$Y_c = a + bX$$

which is called estimated regression line.

Calculation of Regression Coefficients

First the two normal equations are found. This is done from the equation of the straight line:

$$Y = a + bX$$

Hence $\sum Y$ can be expressed as:

$$\sum Y = na + b\sum X \qquad \text{...................(4)}$$

and $\sum XY$ is expressed as:

$$\sum XY = a\sum X + b\sum X^2 \qquad \text{.................(5)}$$

These two equations (4) and (5) are called the normal equations and on these basis the regression coefficients of a and b can be found out directly. The calculation is explained with the help of the information presented in Table 3.3.

Table 3.3: Calculation of Regression Coefficients

	X	*Y*	*XY*	X^2
	0	40	0	0
	1	46	46	1
	2	44	88	4
	3	55	165	9
	4	49	196	16
Total	10	234	495	30

Accordingly the two normal equations are:

$234 = 5a + 10b$

$495 = 10a + 30b$

Out of these two simultaneous equations the values of the regression coefficients calculated are:

$a = 41.4$ and $b = 2.7$

So the regression equation is:

$Y_c = 41.4 + 2.7\,X$

A simple working formula can be derived from the two normal equations of (4) and (5) to find out the values of the regression coefficients. From the normal equation (4) the coefficient 'a' can be derived as under which has been stated earlier:

$$\Sigma Y = na + b\Sigma X \text{ or}$$
$$na + b\Sigma X = \Sigma Y$$
$$na = \Sigma Y - b\Sigma X$$
$$a = \frac{\Sigma Y - b\Sigma X}{n}$$
$$= \frac{\Sigma Y}{n} - \frac{b\Sigma X}{n}$$
$$= \overline{Y} - b\overline{X}$$

The coefficient b can be derived from the two normal equations of (4) and (5) as under:

Multiply the normal equation (4) by X/n as:

$$\frac{\Sigma X}{n}(\Sigma Y = na + b\Sigma X)$$

Therefore:

$$\frac{\Sigma X \Sigma Y}{n} = na\,(\Sigma X / n) + b\Sigma X(\Sigma X / n)$$
$$= a\Sigma X + b\Sigma X\,(\Sigma X/n) \qquad \text{......... (6)}$$

Deduct equation (6) from the second normal equation (5) to eliminate 'a' from the equation to get the value of regression coefficient 'b' as under:

$$\Sigma X\Sigma Y - \Sigma Y\Sigma X/n = (a\Sigma X - a\Sigma X) + (b\Sigma X^2 - b\Sigma X\Sigma X/n)$$
$$= b(\Sigma X - \frac{(\Sigma X)^2}{n}$$

$$\frac{\sum(XY) - \sum X \sum Y / n}{\sum X^2 - (\sum X)^2 / n} = b$$

According to these simple formulas the earlier example worked from the figures of table 3.3 can be calculated easily:

By multiplying both numerator and the denominator by n we can get simple formula further for 'b' as:

$$b = \frac{n\sum(XY) - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} \quad \text{.........} (7)$$

$$b = \frac{5(495) - (10)(234)}{5(30) - (10)^2} = 2.7$$

$$a = \frac{234}{5} - (2.7)\frac{10}{5} = 41.4$$

The least square estimator of b given earlier confirms also to the above formula. The numerator of the formula is $\sum(X - \overline{X})(Y - \overline{Y})$ which can be derived as:

$$\begin{aligned}
\sum(X - \overline{X})(Y - \overline{Y}) &= \sum(XY - Y\overline{X} - X\overline{Y} + X\overline{Y}) \\
&= \sum XY - \overline{X}\sum Y - \overline{Y}\sum X + \sum X\overline{Y} \\
&= \sum XY - n\overline{X}Y - n\overline{X}Y + nX\overline{Y}\,(\sum X = n\overline{X}) \\
&= \sum XY - n\overline{X}Y \\
&= \sum XY - (\sum X \sum Y)/n
\end{aligned}$$

The denominator of the least square estimator can also be explained as:

$$\begin{aligned}
\sum(X - \overline{X})^2 &= \sum(X^2 - 2X\overline{X} + X^2) \\
&= \sum X^2 - 2\sum Xn\overline{X} + n\overline{X}^2 \\
&= \sum X^2 - 2n\overline{X}n\overline{X} + n\overline{X}^2 \\
&= \sum X^2 - 2n\overline{X}^2 + n\overline{X}^2 \\
&= \sum X^2 - 2\overline{X}^2 \\
&= \sum X^2 - (\sum X)^2 / n
\end{aligned}$$

Thus the value of

$$b = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{(\Sigma X - \overline{X})^2}$$

Prediction

Once the two regression estimators, the intercept and the slope, are calculated one can predict Y for any given value of X. This Y is known as estimated Y and denoted as cap $\hat{Y}$ or Y_c.

If the regression function of income (Y) on years of education put up (X) is as follows:

$Y_c = 5000 + 700\ X$

Then for X =10 years of education the estimated value of Y can be computed as:

$Y_c = 5000 + 700\ (10)$

$= \text{Rs.}12{,}000$

This regression model is a predictive model rather than an explanatory model. Of course, the predictive models are not completely distinct from the explanatory models. Commonly a good explanatory model also predicts rarely well. Most of the social scientists are mainly concerned with the identifying of the variables that cause the dependent variables or explanatory models.

Co-efficient of Determination

A researcher wants to know how far the estimated Y is explained by the real data for which the straight line is fitted. In general we know that closer the regression line to the scattered points better the equation fits the data and the error is less. Hence the goodness of fit of a model has to be known and this is indicated by the coefficient of determination R^2.

R^2 can be calculated on the basis of the three deviations of Y series as under:

Total sum of Squared Deviations, TSD = $\sum (Y - \overline{Y})^2$

Regression sum of Squared Deviations, RSD = $\sum (Y_c - \overline{Y})^2$

Error sum of Squared Deviations, ESD = $\sum (Y - Y_c)^2$

The nature of the deviations are graphically shown in Fig. 3.2.

Y– b is the total deviation or TSD

a – b is the deviation explained by the regression line even though the actual point is Y or RSD.

Y – a is the error term of e which can provide the real value of Y as a + bX + e or ESD.

Accordingly the relationship is:

(Y, b) = (a, b) + (Y, a) or

TSD = RSD + ESD

In other words the RSD is the explained part of the component while ESD is the unexplained part. According to the least square method ESD is the minimum in the estimate. Hence the ratio of explained portion to the total deviation is regarded as the goodness of fit or the coefficient of determination R^2.

Hence R^2 = RSD/TSD

The calculation of coefficient of determination is explained with the help of the series given in Table 3.4.

Interpretation of R^2

The co-efficient of determination R^2 indicates the explanatory power of the regression model. It records the proportion of variation

Table 3.4: Calculation of Coefficient of Determination

Height	*Weight*		
	X	*Y*	Y_c
5′3″	0	130	132
5′4″	1	145	142
5′5″	2	150	152
5′6″	3	165	162
5′7″	4	170	172

in the dependent variable explained by the independent variable. The important characteristics are:

According to table 3.4 the mean of Y series is 152 and the regression equation is:

$Y_c = 132 + 10\,X$

TSD		ESD		RSD	
$Y-\overline{Y}$	$(Y-\overline{Y})^2$	$Y-\overline{Y}$	$(Y-\overline{Y})^2$	$Y-\overline{Y}$	$(Y-\overline{Y})^2$
−22	484	−2	4	−20	400
−7	49	3	9	−10	100
−2	4	−2	4	0	0
13	169	3	9	10	100
18	324	−2	4	20	400
	1030		30		1000

TSD = ESD + RSD

1034 = 30 + 1000

$$R^2 = \frac{1000}{1030} = 0.97$$

1. The value of R^2 ranges from +1 to 0. When $R^2 = 1$, the independent variable is completely explains the variance of the dependent variable. The X gets an exact value of Y without any error. In economics or in other social sciences this is not a real case. Here there is no difference between Y and Y_c that means $(Y - Y_c) = 0$.

2. In the other extreme $R^2 = 0$. The points of Y here is so scattered that the regression line becomes horizontal. The independent variable accounts for no variation in the dependent variable. The knowledge of X here is no help in predicting Y, because the two variables are totally independent of each other. Hence R^2 falls in two extremes of +1 and 0. The closer the value of coefficient to 1 the better the fit of the regression line to the points and more variation in Y is explained by X.

3. A very high value of R^2 is essential if predictions are to be accurate, but that does not mean we have a causal explanation for the dependent variable. It merely provides a statistical explanation, it cannot explain what causes income (Y) to vary as a result of changes in education (X). For instance if Y of current income depends on Y_{t-1} and we put as an regression equation, the R^2 may be very high but cannot explain why it happened.

4. The low value of R^2 is though disappointing but it can be informative. It may suggest that they are actually not in linearity but in non-linear form and in close relationship. It may explain that the contribution of X is very small in explaining the Y or non-linear dependency.

Assumptions and Inferences

In a regression analysis a researcher takes a sample analysis and inferes for the population values of the intercept and the slope. For the population equations the Greek letters or English capital letters are used for the coefficients while for the samples small letters are used as under:

Population : $Y_i = A + BX_i + e_i$

Sample : $Y_i = a + bX_i + e_i$

The first assumption is that there is no specification error in the series. That means the functional form of relationship is actually a straight line. Further it is also confirmed that there are no variables improperly included or excluded as causes of the dependent variable. In the income-education relationship if R^2 value is very poor, that means there are other values which influence income other than education and those values have to be identified. Another specification error may be that an independent variable has been included which is irrelevant. Education might not really associated with income which has been taken up for analysis.

The second assumption is that there is no measurement error. When there is a measurement error in case of dependent variable the least square estimates may remain unbiased if the error is random, but if an independent variable is measured with any error the least square estimates will be biased.

The third set of assumptions are involved with the nature of the error term:

(a) It is assumed that its mean is zero. $E(e_i) = 0$. But whatever may be the mean the estimate of the slope will remain unchanged. If this assumption is not met the intercept estimate will be biased. If one is not interested about the intercept this assumption may be ignored.

(b) *Homosedasticity:* The variance of the error term is constant for all values of *x*. This means variance in prediction errors is more or less constant across the values of *x*. That is the points snuggle in a band of equal width above and below the regression line. But on the other hand if the points fanned out from the regression line as the value of x increased this assumption will not hold good. That will be the condition of heterosedasticity. And for this, weighted least square approach has to be adopted.

(c) *Auto-correlation:* This assumption says that the error terms are uncorrelated. This means the errors corresponding to an observation is not correlated with any other errors for other observations. When auto-correlation is present the least square parameter estimates are still unbiased but in this case significant tests and confidence intervals are not valid. Auto-correlation more frequently appears with the time series variables than the cross-sectional variables.

(d) It is assumed that the independent variable is uncorrelated with the error term. If the observed *x* variable is related to the error term, the least square estimate of the parameters will be biased. If they are correlated the missing explanatory data has to be incorporated. But if it cannot be incorporated it has to be assumed that it is uncorrelated.

(e) The last assumption is that the error terms are normally distributed. The distribution of *y* and *e* are the same, that is, normally distributed except in the differences in the mean. As one knows that a normal distribution is a bell shaped curve, with 95 per cent of the observations falling within two standard deviations plus or minus, of the mean. A skewness test can be made of the frequency distribution on the formula below:

$$\text{Skewness} = \frac{\Sigma\left(\dfrac{Y_i - \overline{Y}}{S_y}\right)^3}{n}$$

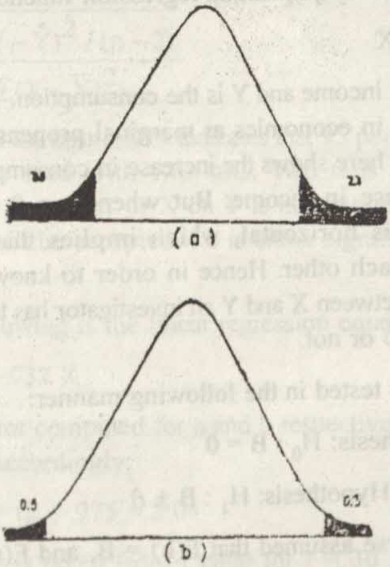


Fig. 3.2. Level of Significance: Two Tailed Test

If the distribution is normal the Skewness = 0, or near about zero.

There are controversies about these assumptions. One view is that the parameters are not meaningfully influenced by the violation of the assumptions. But others like Bibby (1977) believed that the violation of these assumptions would make the regression results almost useless. Of these assumptions the normality assumption can be given up when the sample size is large one because of the operation of the central limit theorem.

Tests of Significance

On the basis that X and Y assumed to be related and they are related in the form of a straight line the regression analysis has been carried on. But then the main question comes whether they are really related? If they are related, the regression coefficient B will have a positive value, if they are directly related and a negative value if they are inversely related. But if they are not at all related B = 0.

For example the population regression function is

$$Y = A + BX$$

in which X is the income and Y is the consumption. This relationship generally known in economics as marginal propensity to consume. The coefficient B here shows the increase in consumption when there is an unit increase in income. But when B = 0, the population regression line is horizontal, which implies that X and Y are independent to each other. Hence in order to know whether there is a relationship between X and Y an investigator has to check whether B is equal to zero or not.

This can be tested in the following manner:

Null Hypothesis: $H_0 : B = 0$

Alternative Hypothesis: $H_1 : B \pm \partial$

It is of course assumed that E(b) = B, and E(a) = A, and the sample distribution of a and b are both normally distributed.

For the testing of these hypotheses one has to select:

1. *Whether the test is one tailed or two tailed.* In a normal curve if both the end are taken for testing it is known as two tailed, and if one side is taken it is known as one tailed test.

2. The confidence interval has to be specified. The widely used confidence intervals are 95 per cent and 99 per cent. In other words, statistical significance at .05 or .01 level.

In testing B = 0, if the value of 0 does not fall within this interval the null hypothesis is rejected and the alternative hypothesis that B ± 0 is accepted. This means the slope estimate b is significantly different from zero.

A two tailed 95 per cent confidence interval model is:

$(b \pm t_{n-2}\ ;\ .975\ s_b)$

The term s_b is called the standard error. It is an estimate of the standard deviation of the slope estimate b. The formula for calculation of this is:

$$S_b = \frac{\sqrt{\Sigma(Y-\hat{Y})^2 / (n-2)}}{\Sigma(X-\overline{X})^2}$$

In the model the figure .975 indicates that 95 per cent confidence interval has been used with two tails. Two tails means that the hypothesis about the effect of X on Y is not one directional, or in other words H_1 will be accepted if b is either significantly negative or positive.

Say the following is the linear regression equation estimated:

$Y = 5078 + 732\ X$

Standard error computed for a and b respectively are 1498 and 118 for n = 32. Accordingly:

$t_{n-2}\ ;\ .975 = t_{30}\ ;\ .975 = 2.04$

The t value is taken from t table for t = 30 and under .025 column, as under two tail model one tail is half of .05. Therefore two-tailed 95 per cent confidence interval for B is

$(b \pm t_{n-2}\ ;\ .975.\ s_b) = 732 \pm 2.04\ (118) = 732 \pm 241$

This shows that the probability is .95 the value of the population slope B is between 491 (732 – 241) and 973 (732 + 241). Since the value zero does not fall within the interval, the null hypothesis is rejected. Hence it can be concluded that b is significantly different from zero at .05 per cent level. Similarly the intercept a can be tested:

$$H_0 : a = 0$$

$$(a \pm t_{n-2} ; .975. s_b) = 5078 - 2.04 (1498)$$

$$= 5078 \pm 3056$$

This shows that 'a' ranges between 2022 and 8134, and hence zero is not in between the two, so the null hypothesis has to be rejected. The intercept estimate 'a' is significant at .05 level. This means the possibility that 'a' cuts the origin of the graph is rejected.

In the above test the null hypothesis is rejected with the probability of 95 per cent, but still probability of 5 per cent that may be wrong. In fact if the null hypothesis is correct, but it is rejected, the Type 1 error is committed. In order to avoid this the same test can be made with confidence interval of 99 per cent that is, probability of 99 per cent for more accuracy. In this test only the t value have to be changed, the t value under column .005 in the t-table for t = 30 is 2.75. Accordingly the values of 'a' and 'b' are:

$$(b \pm t = 30; .995.s_b) = 732 \pm 2.75 (118)$$

$$= \text{Ranges between } 407.5 \text{ to } 1056.5$$

$$(a + t = 30; .995.s) = 5075 + 2.75 (1498)$$

$$= \text{ranges between } 955.5 \text{ to } 9194.5$$

This confidence interval confirms that the null hypothesis is not true and there is no Type 1 error if it is rejected, because the confidence interval does contain value zero.

These tests are shown graphically for the above interpretation in Fig. 3.2. In the two t curves the shaded area shows the level of significance. In the graph (a) the confidence level is 95 per cent, that means the shaded area shows only 5 per cent distributed in each of the two tails at 2.5 per cent. The left tail is negative while the right tail is positive with a t value of 2.04. Similarly in the graph (b) the shaded areas in each tail indicates .5 per cent which is a test for 99 per cent of confidence interval. The t value is 2.75 up to the shaded area of both the sides from the origin which indicates time of standard error distance from the b = 0, either direction positive or negative.

One Tailed Tests

In many cases of linear relationships in social sciences the slope of the linear curve cannot be negative. In case of education-income regression model explained earlier, it is a positive one, here negative means additional education decreases income which is absurd. Hence there is no necessity of a two tailed test hoping for both negative and positive. In this case it is sufficient if one assumes positive slope only, one-tailed test is sufficient to know the relationship.

The hypotheses will be:

$H_0 : B = 0$

$H_1 : B > 0$

We can test the alternative hypothesis in order to reject the null hypothesis:

$$B > b - t_{n-2}; .95\ s_b) = 732 - 1.70\ (118) = 732 - 201$$
$$= 531$$

This shows that b is positive at 95 per cent confidence and hence the null hypothesis is rejected.

A Quick Test

As one knows that the confidence interval will not be zero hence it can be stated:

1. If b is positive $(b - t_{n-2}; .975\ s_b) > 0$
2. If b is negative $(b + t_{n-2}; .975\ s_b) < 0$

The above two can be restated as:

$\frac{b}{s_b} > t_{n-2}; .975$ when b is positive

$\frac{b}{s_b} < t_{n-2}; .975$ when b is negative

A simple formula can be stated for the rejection of the null hypothesis as:

$$\left|\frac{b}{s_b}\right| > t_{n-2};\ 975$$

This means the absolute value of the parameter estimate b, divided by its standard error s_b when exceeds t distribution value of t_{n-2}; .975 the null hypothesis is rejected. It can be seen from the t table that the infinity value of t at 5 per cent two tail level is around 2 (1.96 to be exact). Hence the rule of the thumb can be stated that any value b/s_b which exceeds 2 the null hypothesis is rejected.

$$\left|\frac{b}{s_b}\right| > 2$$

In our income-education example:

$b/s_b = 732/118 = 6.20$

$a/s_b = 5078/1498 = 3.39$

These two t ratios are exceeding 2, hence we can conclude that 'a' and 'b' are statistically significant at .05 level.

When Statistically Not Significant

When b is not statistically significant one can conclude that X is not related to Y or X is not a cause of Y. But this immediate conclusion may not be true. Even though X is in fact related to Y, but it may not show significant statistically because of several reasons.

1. ***Inadequate Sample Size:*** When sample size is increased a given coefficient is more likely to be found significant. Hence if actual significance has to be found the researcher has to select a larger sample, because even though there is a relationship a smaller sample may not reveal it.

2. ***Type II Error:*** When on the basis of this the null hypothesis is rejected which is in fact is true the Type II error is committed. Significant tests can be conducted on any value between 0 and 1. But it is convention among the social scientists, as stated earlier, to assume the statistical levels at .01 or .05. Suppose at the .01 level it is found that estimate is not significant so it is rejected. But if estimates tests is made at .05 level it is

found significant. In the earlier .01 level since it has been rejected which in fact significant the type II error is committed.

3. *Specification error:* In another occasion b may not be significant because of wrong assumption with regard to the relationship of the two variables. The relationship may be curvilinear but it is wrongly assumed as linear, hence b is not significant.

4. *Restricted Variance in X:* Restricted Variance of the X series, the independent variable may be cause of the b being not significant. Since the significant test is mainly based on standard error of estimate as given below:

$$S_b = \frac{\sqrt{\Sigma(Y-\hat{Y})^2/(n-2)}}{\Sigma(X-\overline{X})^2}$$

The above ratio shows that when the dispersion of X about it mean decreases the value of the denominator decreases, and increases the standard error of b. A large standard error makes the statistical significance more difficult to achieve. This means b may not be significant because there is too little variation in X. Hence the researcher has to seek more extreme observations of X, before rejecting a hypothesis.

Prediction Error of Y

In regression analysis the difference between the observed values and the estimated values of Y is known as prediction error or standard error of estimate of Y. This provides a short of average error in predicting Y. Utilising t distribution, approximates 2, at 95 per cent confidence interval Y value can be known as:

$$Y = \hat{Y} \pm 2\, s_e$$

s_e is the standard error estimate as per the following formula:

$$S_e = \frac{\sqrt{\Sigma(Y_i - Y_t)^2}}{n-2}$$

As per our earlier example:

$$Y_c = 5078 + 732\ (10) \text{ for } X = 10$$

$$= 12{,}398$$

S_e calculated for the same series is 2855. Hence the Y value ranges between the two values.

12,398 + 2 (2855) = 18,108

12,398 – 2 (2855) = 6,688

Therefore the real Y values are ranged between 18,108 to 6,688 and hence the prediction is very accurate. In order to predict the value of Y_e more accurately the R^2 value should be much higher, in other words the prediction error would reduce only when the coefficient of determination is large.

Multiple Linear Regression

So far, the regression analysis involved only two sets of variables, of which one is the independent variable, hence it is called bivariate regression model or simply simple regression model. The statistical technique involving two or more independent variables or causal variables, is called multiple regression model. The Multiple Linear Regression model MLR is regarded as more appropriate for all practical purposes than the simple regression model.

1. Multiple regression offers a fuller explanation of the dependent variable. But in simple regression only a single cause is taken which is far from reality especially in case of economic analysis and other social sciences.

2. In the multiple regression the effect of the particular independent variable is made more certain. Because distoring influence of other independent variables is removed from the model.

In the past the main problem in analysing multiple regression was the difficulties of its calculation. When a model of multiple regression involves number of independent variables it was a tedious task, which required special computational methods to calculate the regression coefficients and to conduct the significant tests. But today with the help of computers it can be easily calculated.

The Linear Regression Function

In the general multiple regression equation the dependent variable seen as a linear function of more than one independent variable. The equation can be stated as:

$$Y = a_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots b_kX_k + e$$

The subscripts indicate the number of the independent variables. The three variable model, with two independent variables, which will be analysed here can be stated as:

$$Y = a_0 + b_1X_1 + b_2X_2 + e$$

This shows that Y is determined by X_1 and X_2 and e the error term. The principles of least squares is applied here as in case of simple regression model which minimises the sum of the squares of the prediction error: $\Sigma(Y - \hat{Y})$. Hence the three variable least square equation is:

$$Y = a_0 + b_1X_1 + b_2X_2$$

Alike the simple regression model, this three variable regression model cannot be shown graphically. Here one must imagine fitting a plane to a three dimensional scatter points. If there are more than two independent variables in the model it cannot be presented in graphs. For k independent variables it requires k-dimensional hyperplane to a (k+1) dimensional scatter.

Estimate of the Coefficients

The intercept a: The interpretation is same as in case of the bivariate regression model. It is the average value of Y when each independent variable equals zero.

The Slopes b_1 and b_2: The b's here indicate the average change in Y associated with a unit change in X_1 when the other independent variable X_2 is held constant. By keeping the other independent variable constant, its distorting influence is eliminated. Since here there are two slopes which explains the behaviour of Y, each one is called a partial slope. Hence b_1, b_2 are the partial slopes or partial regression coefficients.

In estimating the partial slopes it is assumed that X_1 and X_2 are related or $r_{12} = 0$. That means each independent variable is accounted for at least in part by the other independent variable. That is:

$$X_1 = c_1 + c_2X_2 + u$$

$$\hat{X}_1 = c_1 + c_2X_2$$

Hence $X_1 = \hat{X}_1 + u$

or $u = X_1 - \hat{X}_1$

The u represents a part of X_1 which is completely separated from X_2. On the same analysis it can be proved that:

$$Y = d_1 + d_2X_2 + v$$

$$\hat{Y} = d_1 + d_2X_2$$

$$Y = \hat{Y} + v$$

$$V = Y - \hat{Y}$$

This v is the part of Y which is unrelated to X_2. These two error terms components are joined together to explain the partial coefficient b_1

$$b_1 = \frac{\Sigma(u_1)(v)}{\Sigma u_1{}^2}$$

$$= \frac{\Sigma(X_1 - \hat{X}_1)(Y - \hat{Y})}{\Sigma(X_1 - \hat{X}_1)^2}$$

In other words b_2 is determined by X_1 and Y values that have been freed of any linear influence from X_2. Similarly b_2 can be calculated in the formula:

$$b_2 = \frac{\Sigma(u_2)(v)}{\Sigma(u_2{}^2)}$$

Computation of the Coefficients

The b's can be calculated on the basis of the three normal equations derived from

$$Y = a + b_1X_1 + b_2X_2$$

The normal equations are:

Multiplied by Σ : $\Sigma Y = na + b_1\Sigma X_1 + b_2\Sigma X_2$ ..... (6)

Multiplied by ΣX_1 : $\Sigma X_1Y = \Sigma_{X1} + b_1\Sigma X^2{}_1 + b_2\Sigma X_1X_2$ ..... (7)

Multiplied by ΣX_2 : $\Sigma X_2Y = a\Sigma_{X2} + b_1\Sigma X_1X_2 + b_2\Sigma X_2{}^2$ ..... (8)

From these three simultaneous equations the value of b_1 and b_2 can be calculated. This can be further simplified by shifting the origin from (0, 0, 0) to $(\overline{Y}, \overline{X}, \overline{X}_2)$

When the first of the normal equation is divided by n the intercept 'a' can be derived easily:

$$\frac{\Sigma Y}{n} = \frac{na}{n} + \frac{b_1 \Sigma X_1}{n} + \frac{b_2 \Sigma X_2}{n}$$

$$\overline{Y} = a + b_1 \overline{X}_1 + b_2 \overline{X}_2$$

$$a = \overline{Y} - b_1 \overline{X}_1 - b_2 \overline{X}$$

Deducting the above equation from the main regression equation we get:

$$Y = a + b_1 X_1 + b_2 X_2$$

$$\overline{Y} = a + b_1 \overline{X}_1 + b_2 \overline{X}_2$$

$$(Y - \overline{Y}) = 0 + b_1 (X_1 - \overline{X}_1) + b(X_2 - \overline{X}_2)$$

$$(Y - \overline{Y}) = 0 + b_1 (X_1 - \overline{X}_1) + b(X_2 - \overline{X}_2)$$

In deviation notifications:

$$y = (Y - \overline{Y});\ x = (x - \overline{X})$$

Hence the equation becomes:

$$y = 0 + b_1 x_1 + b_2 x_2$$

On the basis of this the three normal equations can be given as

$$y = n.0 + b_1 x_1 + b_2 x_2$$

$$\Sigma x_1 Y = 0.\Sigma x_1 + b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2$$

$$\Sigma x_2 Y = 0.\Sigma x_2 + b_1 \Sigma x_1 x_2 + b_2^2 x_2^2$$

As the first terms of the each equation are zeros the last two normal equations can be written as:

$$b_1 \Sigma x^2_1 + b_2 \Sigma x_1 \Sigma x_2 = \Sigma x_1 Y$$

$$b_1 \Sigma x_1 \Sigma x_2 + b_2 \Sigma x^2_2 = \Sigma x_2 Y$$

The working formula can be stated as:

$$\Sigma x_1^2 = \Sigma (X_1 - \overline{X}_1)^2 = \Sigma X_1^2 - n (\overline{X}_1)^2$$

$$\Sigma x_1 x_2 = \Sigma (X_1 - \overline{X}_1)(X_2 - \overline{X}_2) = \Sigma X_1 X_2 - n (\overline{X}_1)(\overline{X}_2)$$

An example is worked out here for the calculation of multiple regression coefficients with the help of Table 3.5 (a) and (b)

Table 3.5 (a): Calculation of Multiple Regression Coefficient

	Y	X_1	X_2
	7	4	1
	12	7	2
	17	9	5
	20	12	8
Total	56	32	16
Mean	14	8	4

All the required values for the computation of the regression coefficients can be had from the following Table 3.5.

Table 3.5 (b)

	X_1^2	X_1X_2	X_2^2	X_1Y	X_2Y	Y^2
	16	4	1	28	7	49
	49	14	4	84	24	144
	81	45	25	153	85	289
	144	96	64	240	160	400
Total	290	159	94	505	276	882

$$\sum x_1^2 = 290 - 4\,(64) = 34$$

$$\sum x_1x_2 = 159 - 4\,(8)\,(4) = 31$$

$$\sum x_2^2 = 94 - 4\,(16) = 30$$

$$\sum x_1y = 505 - 4\,(8)\,(14) = 57$$

$$\sum x_2y = 276 - 4\,(4)\,(14) = 52$$

Thus the normal equations are:

$$34b_1 + 31b_2 = 57$$

$$31b_1 + 30b_2 = 52$$

and solving b_1 and b_2 we find:

$$b_1 = 1\frac{39}{59}; b_2 - \frac{1}{59}$$

Accordingly the value of intercept 'a' is:

$$\begin{aligned} a &= \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2 \\ &= 14 - (98/59)8 - (1/59)4 \\ &= 38/59 \end{aligned}$$

Hence the regression equation is:

$$Y_c = \frac{38}{59} + \frac{98}{59}X_1 + \frac{1}{95}X_2$$

Significance Tests

The partial slope estimates b_1 and b_2 can be tested whether they are significantly different from zero or not; that is, the null hypothesis has to be tested. The same procedures can be followed as in case of the bivariate regression.

$$(b_1 \pm t_{n-3} \, ; .975 \, s_b)$$

This is same thing as in case of the bivariate test except that the degree of freedom is n – 3, as there are three variables in the model. The principle for the degree of freedom is, the degree of freedom of t variable is equal to (n – k – 1), where n is the number of sample size and k = number of independent variables.

Alternatively if b_1 and b_2 divided by their respective standard errors is more than two, it can be taken as significant and null hypothesis can be rejected according to the rule of the thumb as:

$$b_1/sb_1 > 2 \text{ and } b_2/sb_2 > 2$$

F-Test

In the multiple regression the testing is of a complicated affair. The test here is simply for the individual partial slopes of the independent variables but the joint effect of all the independent variables on the dependent variable has to be tested to find out how far it is significantly different from zero.

The null hypothesis should be:

$H_0 : B_1 = B_2 = B_3 = \ldots\ldots\ldots\ldots\ldots B_k = 0.$

For this the appropriate test statistic is F which can be found with the help of the following formula:

$$F = \frac{R^2 / k}{(1 - R^2 / (n - k - 1)}$$

In this equation R^2 is the squared multiple correlation coefficient for the full regression model. The constant k is the number of independent variables in the model.

The null hypothesis is rejected when the actual value of F from the model calculated according to the above formula is greater than the critical value at the level of significance found from the F table.

Suppose that R^2 of a regression model with 2 independent variables is .45 and n = 28. According to the above formula F value is:

$$F = \frac{.45/2}{(1-.45)/25} = \frac{.45/2}{.55/25}$$

$$= \frac{.45}{.55} \times \frac{25}{2}$$

$$= 10.23$$

In the above formula k is 2 that is the number of independent variable which is denoted as n_1. The number 25 in the denominator is the degrees of freedom, since the degree of freedom is n – k – 1 and n being 28. This is denoted as n_2. Now find out the critical value of F at the significant level of .05. This can be found out from the F table given in any statistical text for n = 2; and $n_2 = 25$ which is 3.28.

Calculated F		Critical F
10.23.	>	3.28

The calculated value is greater than the critical value at .05 level hence the null hypothesis is rejected that means it is concluded that there is significant relation among the variables.

In another example when $R^2 = .23$; $n = 16$ and with 3 independent variables $n = 3$ and $n_2 = 12$ the actual value in F statistic is less than the critical value at .05 level as shown under;

Calculated F		Critical F
1.19	<	3.49

In this case the null hypothesis is accepted, that means it is confirmed that the relationship is not significant at .05 level.

Multi-collinearity

In the multiple regression, apart from all the bivariate regression assumptions, there is one additional assumption that there is no perfect multicollinearity among the independent variables. In other words non of the independent variable is perfectly correlated with another independent variable or the linear combination of other independent variables.

In the multicollinearity it is assumed that perfect multicollinearity exists between two independent variables X_1 and X_2 if they are

$$X_2 = c_0 + c_1X_1$$

When there is perfect collinearity the $R^2 = 1$. In this case unique solution for the least square parameter estimate are not possible; so the partial regression coefficients are not possible for calculation. But in practice perfect multicollinearity is not generally found. There may be high multicollinearity but not perfect. With regards to the multicollinearity three things can be noted down:

1. The multicollinearity is a problem referring to the correlated independent variables in a specific sample data and not in the overall population. The social scientists may collect data where the independent variables are not strongly correlated.
2. A high multicollinearity present in a study does not violates the assumptions of the regression. The estimates are regarded still unbiased.

3. It is a fact that multicollinearity exists in every situation, hence it is not the question of 'exists' or 'does not exists' but more particularly the degree of multicollinearity. A low degree of multicollinearity is of not very much concern but a high degree of multicollinearity poses certain problems.

As it is stated earlier the s_b the standard error is very much influenced by the correlation between the independent variables. A larger standard error is a major consequence of multicollinearity but it is not the sole cause. Every high standard error cannot be interpreted as the cause of the multicollinearity. There are two other causes of high standard error: [1] it may be due to small sample for estimation or [2] due to the variables with small variances in the sample.

Detection of Multicollinearity

The following are some of the simple methods of detecting the existence of multicollinearity in a model:

1. A multicollinearity is suspected when one of the 't' ratios of the regression coefficients for the independent variables is sufficiently large to indicate statistical significance at the .05 level, while F statistic for the full model is significant.
2. Another method of testing the multicollinearity is the laying out of the bivariate correlations in the matrix form. In the inspection if any of the correlations are not exceeding a predefined cutoff value say .80 then it can be concluded that there is no high degree of multicollinearity. But this is not a foolproof method. A variable may not be any high correlation with any single variable but may be very significant taking several variables together. Further it is very difficult to set a cutoff value for the detection.
3. By regression one independent variable on the rest of the other independent variables in the model and finding R^2 the multicollinearity can be tested. If any of the R^2 value is close to 1.0 then it can be concluded that there is high degree of multicollinearity present in the model. This test is superior to the bivariate correlation coefficient but still it has not overcome

the difficulty in defining how high the correlation must be before one takes concern about the presence of multicollinearity.

How to Avoid Multicollinearity?

If a researcher finds strong multicollinearity in a model what has to be done to get over the problem? There are two alternatives for a researcher:

1. One of the best solution is to obtain additional information. This means increasing the size of the sample n. When the sample size is increased, keeping all other factors in the formula of calculating standard error constant, the standard error decreases and can offset the effect of multicollinearity. But increases in 'n' may not result in decrease in standard error. And some times it is not possible to get additional information to increase 'n'.
2. The second course left for a researcher to overcome the problem of multicollinearity is use of regression coefficients for the independent variables which are known to the researcher from his previous study or from the studies of similar nature from others. If one knows that the population coefficients of two independent variables are related in a particular ratio then the extreme multicollinearity of the model can be overcome. For instance the model is:

$$Y = A + B_1X_1 + B_2X_2 + E$$

Where the independent variables have sample correlation near 1.0 but previous research tells us that

$$B_1/B_2 = C$$

Hence $CB_2 = B_1$; so we can substitute CB_2 for B_1. Hence the new equation would be:

$$Y = A + CB_2X_1 + B_2X_2 + E$$

which can be written as

$$Y = A + B_2X_3 + E$$

where $X_3 = CX_1 + X_2$, and accordingly B_2 and A has to be estimated and one B_2 is known as per the ratio relations B_1 can be calculated.

Of course if one of the regression coefficients is directly known from the earlier research it can be replaced outright in the equation for the estimation of multicollinearity.

If non of the above course is available one has to accept the multicollinearity and as such the analysis may be abandoned.

Dummy Variables

In most of the statistical analysis generally the analysis is made at the interval level, that is, numerical values are assigned to the variables for the purpose of comparison precisely and accurately. But all variables are not measured at a level which allows precise comparative values. In regression analysis non-interval variables can be introduced through the employment of dummy variables.

Many non-interval variables can be considered dichotomies or measured in binary values. For example, sex can be treated as male/female, race as black/white or marital status as married/unmarried. In this case the independent variables can take only two values $X = 0$, $X = 1$, female and male respectively.

In the regression model $Y = a + bX$; if $X = 1$ then

$$Y = a + b$$

and if $X = 0$ then

$$Y = a$$

The slope estimate b indicates the difference between the mean incomes of male and female. The t ratio of b measures its statistical significance.

The non-interval variables with multiple categories can be divided into two groups: (1) ordinal level and (2) nominal level. In the first case the variables can be ordered in terms of amounts but cannot be with numeric precision. Attitudal variables are of this kind, where opinions are expressed in order of preference. If a question is asked to a beneficiary whether he is satisfied with the programme; he can express the answer as: "very good", "good" or "satisfactory." Ordinal variables only admit a ranking from less to more or vice versa.

Categories of nominal variables cannot be ordered. For example a persons religion: Hindu/Muslim/Christianity or nature of profession one adopts cannot be ordered or numerically expressed. That means the personal attributes cannot be meaningfully ranked.

The ordinal and personal attributes can be used in a multiple regression in terms of a dummy variable. For example, let the model is:

$$Y = a + b_1X_1 + b_2X_2 + e$$

X_1 : very good = 1 or otherwise = 0

X_2 : good = 1 or otherwise = 0

X_3 : satisfactory

This third dummy variable need not be inserted in the multiple regression because if the first two variables are known the last one is automatically known. Again if all the three are inserted in the model it will describe a perfect multicollinearity, in which estimation cannot be processed. Hence, the general rule is "when a non-interval variable has G categories, G-1, dummy variables represent the model.

$$Y = a_0 + b_1X_1 = b_2X_2$$

In this case when both the X_1 and X_2 are zeros then Y is equal to the intercept. That is $Y = a_0$

When $X_2 = 0$, then the equation will be:

$$\begin{aligned} Y &= a_0 + b_1\,(1) + b_2\,(0) \\ &= a_0 + b_1 \end{aligned}$$

When $X_1 = 0$, then the equation will be:

$$\begin{aligned} Y &= a_0 + b_1\,(0) + b_2\,(1) \\ &= a_0 + b_2 \end{aligned}$$

In social science researches where the study concern with the ordering of the variables and/or confined to personal opinions the dummy variables are very much helpful.

4

Analysis of Nominal Data

Most of the analysis in social sciences are made by exhibiting the data in two-way cross classification tables. There are many types of variables but broadly they can be classified into quantitative and qualitative variables. Quantitative variables are expressed in terms of real numbers for analysis. Most of the statistical analysis are mainly work out the quantitative variables. Values expressed in terms of kilograms, meters and monetary terms are all quantitative measurements. But in social sciences one more often faces the problem where the variables cannot be measured quantitatively. Hence some form of qualitative assessment have to be made for investigation which depends upon the investigator's arbitrary assignment of values to the variables.

There are two basic types of qualitative variables: Nominal and Ordinal. A nominal variable consists of a set of mutually exclusive categories or states so that each observation is assigned just one of the categories. An ordinal variable is also a set of mutually exclusive state but the ordinal variables states are ordered or ranked in terms of the alternative amounts or degrees of intensity that the states represent.

The data in Table 4.1 consists of a sample of 500 adults' cross-classified by their habit of drinking at home, at bar and at both the places.

The data in the table 4.1 of both the variables, income group and drinking habits, are nominal data and the variables are nominal variables. Each sample individual is assigned to a category, which

Table 4.1: Drinking Habit of People

Drinking Habit	*Income Group*			
	Rich	*Middle*	*Lower*	*Total*
Home	75	15	21	111
Bar	30	74	160	264
Both the places	66	17	42	125
Total	171	106	223	500

is homogeneous with respect to the attribute. In other words they are mutually exclusive and exhaustive too. A nominal variable can be arranged in any particular way according to the need of the investigator without loosing information in it. Nominal scales measure truly discrete phenomena. but in most of the case represent measurement error because the underlying traits are more or less quantitative.

The nominal data or categorical data has three important features which are shown in the table.

1. Each individual is assigned to one and only one class according to a particular trait or attribute.
2. The category labels are simply names that indicate how groups differ from one another.
3. The labels say nothing about the magnitude of the differences among the groups.

Number of Categories

Nominal variables are ordinarily organised in two-way or multiway tables, but there is no principle how many categories one should make. It is believed that in cross classification data number of zeros should be avoided to make the analysis reliable and impressive. In such cases combining categories in order to increase frequencies may solve the problem, but there are two difficulties in this procedure.

1. A nominal variable in part, depends upon the number of categories, other things remaining constant, greater the number of classes the greater the variation. Hence the combining of the categories reduces the significance of the analysis.
2. Combining the number of categories often seriously affects the observed relationships. Relations in dichotomised data make differences in the trichotomised data. Hence the general principle should be to retain as many categories as possible and do not dichotomise without good reasons for it.

Dependent and Independent Variables

Dividing nominal variables into dependent and independent, sometimes make irrelevant and arbitrarily assigned by the investigator. A variable that depends on is caused by, or temporarily follows another variable is a dependent variable. In Table 4.1, the categories of economic groups (high, middle and low income group) are the independent variables while the place of use of alcohols are the dependent variables. Only an assumption is made regarding the nature of the variable, since it is not possible to prove that one variable causes another. Generally the independent variables are presented in the columns as X, and the dependent variables in rows as Y.

Sampling

A sampling of nominal variables can be drawn on simple random basis from the population. The marginal totals (the totals of the rows or columns) are not predetermined by the investigator, it appears in the sample by chance. The number of categories are also not determined in the research design. These samples sometimes are called 'multinominal' samples and the variables generated in this way are called responses.

In some samples when the marginal totals are predetermined these variables are called 'factors'. Responses are usually considered as dependent and the factors as independent variables. Research designs involving only response variables occur most frequently in the sample surveys.

Notations

In cross classifying subjects on the basis of two attributes one creates a 'two dimensional table'. The dimensionality refer to the number of variables. Excluding the marginal totals, the table 4.1 has 3 rows and 3 columns, and it is referred as 3 × 3 or three by three table; which indicates the dimension of the table. In algebraic notation it has I rows and J columns, hence it is a I × J table. The notations of a cross classified table is presented in Table 4.2.

Table 4.2: Notations of a Cross-Classification Table

X : Independent Variable

		1	*2*	*3*	*j*	*J*	*Total*
	1	n_{11}	n_{12}	n_{13}	n_1j	n_1J	n_{1+}
	2	n_{21}	n_{22}	n_{23}	n_2j	n_2J	n_{2+}
Y :	3	n_{31}	n_{32}	n_{33}	n_3j	n_3J	n_{3+}
Dependent	i	n_{i1}	n_{i2}	n_{i3}	n_ij	n_iJ	n_{i+}
Variable	I	nI_1	nI_2	nI_3	nIj	nIJ	nI_+
Total		n_{+1}	n_{+2}	n_{+3}	n_{+j}	n_{+j}	n

As per the matrix notations the small n's represents the cell frequencies of the number of cases in a particular row-column classification. Hence n_{ij} represents the number of individuals in the i^{th} row and j^{th} column and the subscripts i and j can take any value from 1 to I and 1 to J respectively.

The totals at the bottoms and sides constitute the marginal distribution of X and Y. At the end of the first row n_{1+} means sum of all observations in the first row. The plus sign in the subscript indicates that all the entries in the first row has been added over the J columns. In the Table 4.1 : $n_{1+} = 75 + 15 + 21 = 111$ and $n_{+1} = 75 + 30 + 66 = 171$. The n_{+1} is the sum total of the first column of the table. The total number of observations is n, i.e. it is 500 in the said table. This quantity can be found by summing all the n's in the table or by summing the marginal totals of either variables.

Chi-Square Test

The most important analysis of nominal data is to test whether the variables are statistically independent in the population? This question is usually answered with the help of a Chi-square test.

The Chi-square test for independence consists of four parts:

1. The null hypothesis that the variables in the study are statistically independent.
2. Expected frequencies are derived under the assumption that the null hypothesis is true.
3. A comparison is made of these expected values with the corresponding observed frequencies.
4. Finally a judgement is made whether or not the difference between expected and observed frequencies could have arisen by chance.

The Null Hypothesis

The null hypothesis is that in a population the two variables in the cross-classification table are statistically independent. The statistical independence holds good if,

$P_{ij} = P_{i+} . P_{+j}$ for all i, j.

Where P_{ij} is the probability of being in ith category of Y and jth category of X.

P_{i+} is the marginal probability of being in the ith category of Y.

P_{+j} is the marginal probability of being in the jth category of X.

Hence the probability of particular income group and behaviour of drinking for a particular individual is the product of corresponding marginal probabilities.

P (Y_iX_j) denote the probability that an individual is in the ith class of Y and jth class of X in an I × J population cross classification. In the probability matrix of nominal variables the cell probability can be expressed as P_{ij} which is same as P (Y_iX_j)

Hence the marginal probabilities can be stated as:

$P_{i+} = P_{i1} + P_{i2} + P_{i3}$ .... $+ P_{ij}$.

$P_{j+} = P_{1j} + P_{2j} + P_{3j}$ .... $+ P_{ij}$

Expected Value

If the null hypothesis is true what are the expected frequencies in the sample cells? The calculation is based on the logic that if there is no association between income groups and their drinking behaviour the proportion of consumption of alcohol at home should be same in all three categories of income groups. From the sample data given in the Table 4.1, the proportion of home consumption can be calculated as:

111/500 = 0.222

or about 22 per cent exclusively drink at home. On this basis out of 171 of higher income groups individuals it is expected that 22.2 per cent of them would have been consumed at home. This can be calculated as:

0.222 X 171 = 37.96

This is the expected value of cell n_{11} which can be denoted as $\hat{n}_{11}$ in algebraic notation.

Hence the expected cell frequencies can be calculated on the basis of the following formula under the null hypothesis.

$$n_{ij} = \frac{(n_{i+})(n_{+j})}{n}$$

Where the n_{i+} and n_{+j} are the ith and jth row and column marginal frequencies respectively as stated earlier.

Table 4.3: Expected values and Components of Chi-Square Statistics

			X: Income Groups		
			Rich	*Mid.*	*Low.*
	Home	n_{ij}	75	15	21
		n_{ij}	37.96	23.53	49.51
		e_{ij}	36.14	3.09	16.42
Y: Place of Consumption	Bar	n_{ij}	30	74	160
		n_{ij}	90.29	55.97	117.74
		e_{ij}	40.26	5.81	15.17
	Both	n_{ij}	66	17	42
		n_{ij}	42.75	26.50	55.75
		e_{ij}	12.64	3.41	3.39

On the basis of this formula the expected values of each cell data of Table 4.1 are presented in the second row of each nominal variable in Table 4.3. The first row indicates the observed values and the third row represents theoretical Chi-square values.

Comparison of Observed and Expected Values

It can be observed that in some of the cells of Table 4.3 the discrepancy between the observed and expected frequencies are very large while in some others it is less. How to measure the discrepancy among these two values? For the overall comparison of the two values the goodness of fit Chi-square can be adopted. The formula for it is as under:

$$X^2 = \sum_i^I \sum_j^J \frac{n_{ij}^2}{\hat{n}_{ij}} - n$$

On the basis of observed and expected values presented in the Table 4.3, x^2 can be calculated as:

$$x^2 = \left[\frac{(75)^2}{37.96} + \frac{(15)^2}{23.53} + \frac{(20)^2}{49.51} \ldots\ldots + \frac{(42)^2}{55.75}\right] - 500 = 136.33$$

This 136.33 is called the observed Chi-square statistic and by comparing this with a critical Chi-square one has to decide whether or not the observed values arose by chance.

In Table 4.1 since there are 3 dependent variables and also 3 independent variables its degree of freedom or df = 4; calculated on the formula df = (I – 1) (J – 1). Three critical values and three Chi-square values at df = 4 are given below for comparison:

Critical values :	.05	.01	.001
x^2 values :	9.488	13.277	18.465

X^2 values calculated earlier for the data in Table 4.1 of 136.33 is clearly higher to these critical Chi-square values presented above; hence the null hypothesis of independence is rejected. This means the two variables are related but not independent of each other. This Chi-square test tells only that there is a relationship between the two nominal variables, but it provides no other information beyond this.

Relationship Between the Variables

With the help of the following two methods the relationship of the two variables can be known:

1. Analysing the components of Chi-square.
2. Analysing by partitioning the Chi-square.

Components of Chi-square: For each observed and expected value, an individual Chi-square statistic can be computed with the following formula.

$$e_{ij} = \frac{(n_{ij} - n_{ij})^2}{n_{ij}}$$

The third rows of each independent variable in Table 4.3 is calculated on this basis. For

$$e_{11} = \frac{(75-37.96)^2}{37.96} = \frac{1371.962}{37.96} = 36.14$$

About 68 per cent of the total X^2 (136.33) is due to the three cells of Table 4.3 : n_{11}, n_{13} and n_{21}. The table with these individual x^2 values show that the higher income groups in their behaviour at home consumption of wine and also lower income group at home consumption of wine are more significant.

Partitioning Chi Square: In this method a table is divided into various sub-tables. Then each sub-table is analysed as a separate cross classification. They can be further divided into sub-tables each having a Chi-square and degree of freedom ultimately each with one degree of freedom. The original Table 4.1 which is in 3 × 3 dimensions can be divided into 4 sub-tables 2 × 2 dimensions as shown in Table 4.4. The total Chi-square is also divided into parts, each pertaining to a separate table and hypothesis.

Chi-squares of the four partitioned sub-tables provide good information to know the relationships. The results are presented in Table 4.5.

Table 4.4: Partitioning of the Table 4.1

	(a)				(b)		
	R	M			RM	L	
H	75	15	90	H	90	21	111
B	30	74	104	B	104	160	264
	105	89	194		194	181	375
	(c)				(d)		
	R	M			RM	L	
HB	105	89	194	HB	194	181	375
Both	66	17	83	Both	83	42	115
	171	106	277		277	223	490

Table 4.5: Chi-Square of the Partitioned Tables

	Relation	X^2	*df*	*Sig. Level*	*Critical value*
1.	R/M vs. H/Bar	57.77	1	.001	10.127
2.	RM/L v. H/Bar	43.26	1	.001	
3.	R/M v. HB/Both	15.87	1	.001	
4.	RM/L v. HB/Both	19.76	1	.001	
	Total	136.65	4	.001	18.465

As it can be seen from Table 4.5 that the first two sub-tables together constitutes the Chi-square value of 101.03 which is 74 per cent of the total Chi-square.

The four partitions shown in the Table 4.5, are not the only category of partition, partitioning can be done may different in many different sub-tables. While partitioning the main table into sub-tables the following rules of the partitioning have to be observed:

1. Each frequency in the original table must appear only once in any of the sub-tables as a frequency. The marginal frequencies must also appear once.

2. The marginal totals in the main or sub-tables may appear in the cells of the sub-tables.
3. The total of frequencies must appear once.

Measures of Association

The goodness of fit Chi-square test tells only that whether the variables are statistically independent or not. If they are not independent, i.e. they are related, the degree of relationship between the nominal variables are not known. Hence a measure of association is necessary. A measure of association is a numerical index summarising the strength or degree of relationship in a two dimensional cross classification. According to the nature of tables the association analysis can be divided into two:

1. First is the 2 × 2 table which is the simplest.
2. Second is the I × J table where I ≠ J, that means it is not a square table.

Measures in 2 × 2 Tables

Association measures among the nominal variables in 2 × 2 tables or square tables can be calculated in any of the five methods given below:

A. **Percentage Method:** This is the most simple method of measure of association. A difference in percentage or proportions can be interpreted as a regression coefficient between the two dichotomous variables. An example is shown in Table 4.6.

Table 4.6: Association Measure in Percentage Method

	X	
	32 (40%)	36 (45%)
Y	48 (60%)	44 (55%)
Total	80 (100)	80 (100)

The table shows that there is 5 per cent difference in the two columns of X. In other words a change in one unit of X produces a change of .05 in Y. Of course it is assumed that the independent and the dependent variables are specified unambiguously. In order to get a reasonable estimate the each category of X must have large values but not less than 20.

B. **Cross Product Ratio:** The cross product ratio (CPR) or sometimes called odd ratio is another method for finding relationship between the two nominal variables. In this analysis the marginal totals and the total number of data n is regarded as fixed.

Observe the sub-table 'a' of the Table 4.4.

	R	M	
H	75	15	90
B	30	74	104
	105	89	194

The odds ratio for drinking for Rich : 75/30 = 2.50

The odds ratio for drinking for Middle : 15/74 = 0.20

This shows that the drinking habits of rich are higher to the middle income group. The odds of the two groups are different but the CPR provides a single indicator of association for better measure between the two nominal variables.

The formula for the CPR is:

$$\mathrm{CPR} = \frac{n_{11}n_{22}}{n_{21}n_{12}}$$

Accordingly the CPR for the above Table is:

$$\mathrm{CPR} = \frac{(75)(74)}{(15)(30)} = 12.33$$

The main characteristic of the CPR is in ranges from 0 to infinity and when the value of CPR is

1. Above 1.00 to infinity the association is taken as positive as in case of the above table.

2. Just 1.00 it indicates nil association or statistical independence between the variables.
3. Below 1.00 it implies a negative association. The CPR of the sub-table d of Table 4.4 has negative association.

$$\text{CPR} = \frac{(194)(42)}{(181)(83)} = 0.542$$

Yule's Q: Yule's Q is the function of the CPR but unlike CPR, the Q lies between –1.0 and +1.0, where 0 implies statistical independence. The simple formula for the estimated $\hat{Q}$ is:

$$\hat{Q} = \frac{\text{CPR} - 1}{\text{CPR} + 1}$$

Accordingly for sub-table a of Table 4.4 Yule's Q can be calculated as under:

$$\hat{Q} = \frac{12.33 - 1}{12.33 + 1} = \frac{11.33}{13.33} = 0.849$$

Many investigators feel that Yule's Q overstates the strength of an association. In any event value close to 1.00 indicates a strong relationship.

C. **ϕ^2 Measures based on Chi square:** One defect of goodness of fit chi square as a measure of association is that its numerical magnitude depends partly on the size of the sample. Dividing chi square by n corrects this, and leads to a popular measure of association phi square ϕ^2.

$\phi^2 = X^2/n$

In the first sub-table of Table 4.5 $X^2 = 57.77$ accordingly ϕ^2 can be calculated as

$\phi^2 = 57.77/194 = .298$

ϕ^2 values vary between 0 and 1.0 and 0 indicates that the variables are statistically independent. In this measure of association the marginal variations in X or Y affects the magnitude of ϕ^2. Greater the imbalance in the marginal distribution the lower is the values. In the sub-table (d) of Table 4.5; $\phi^2 = 0.04$. This is because of the fact that there is high marginal distribution and the total of n is also very high of 490. This is a symmetric index, and does depend upon the variable which is considered dependent. This ϕ^2 is equivalent to r^2 the product moment correlation coefficient.

D. Correlation Coefficient r: Pearson's Product Moment correlation is a good measure of association. In a 2 × 2 table it can be calculated in the following formula:

$$r = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1+}n_{2+}n_{+1}n_{+2}}}$$

Based on this formula r can be found out for Table 4.4 (a) as:

$$r = \frac{(75)\,(74) - (30)\,(15)}{\sqrt{(105)\,(89)\,(104)\,(90)}} = 0.376$$

The r^2 is commonly interpreted as the percentage variation in the dependent variable that explained by the independent variable. Here $r^2 = 0.141$; that is, it explains only 14 per cent. The correlation coefficient is sensitive to marginal distribution but it is invariant under interchange of both rows and columns. As r varies from –1 to +1; when it is zero the nominal variables are independent of each other.

Association for I × J Tables

More often one has to analyse tables with more than 2 × 2 tables where Is and Js exceed 2, whether square or otherwise. The same analysis of 2 × 2 tables can be extended to the I × J tables with an object of reducing influence marginal distribution or table dimensions.

Phi square is calculated based on the chi square in the 2 × 2 tables, but it cannot be used in a I × J table because there is no upper bound. Hence a number of normed variations of ϕ^2 has been proposed to eliminate the above defects. But this measure has one difficulty is that it cannot frequently attain maximum; and the values lying between 0 and 1 are hard to interpret. Two types of measures using ϕ^2 as basis are: (a) Contingency coefficient C and (b) Tichuprow's T value.

Contingency Coefficients: The contingency coefficient can be calculated on the basis of Phi square or Chi square as under:

$$\therefore \quad C = \sqrt{\frac{\phi^2}{\phi^2 + 1}} \qquad \sqrt{\frac{X^2}{X^2 + n}}$$

Contingent coefficient for table presented in Table 4.1 can be calculated as under:

$$C = \sqrt{\frac{136.33}{136.33 + 500}} = 0.46$$

When C equals zero the variables in the table are statistically independent. Here one difficulty is that C does not always reach 1.0 even when the variables are in perfect association. In the square tables (i.e. I = J) its maximum value is $\sqrt{(I-1)/I}$; hence in this case one can have an adjusted C. Accordingly an adjusted C is calculated as

$$C_{adj} = \frac{C}{C_{max}}$$

C_{max} is the maximum value of C for a particular table. For the Table 4.1 the maximum value is then

$$C_{max} = \sqrt{(3 - i)/3} = .816$$

Accordingly C = (.46)/(.816) = .56

For the symmetric tables this type of adjustments are less feasible hence some investigators show that C cannot be used for the tables smaller than 5 × 5.

Tichuprow's T value

The T value varies between 0 and 1.0 but can reach maximum only in the square matrices. When I ≠ J, T is always less than 1.0. The formula for the calculation of T is:

$$T = \sqrt{\frac{\phi^2}{(I-1)(J-1)}} = \sqrt{\frac{\chi^2}{n\sqrt{(I-1)(J-1)}}}$$

Accordingly T value for the Table 4.1 is

$$T = \frac{136.33}{500\sqrt{(2)\,(2)}} = .369$$

5

Analysis of Ordinal Data

An ordinal variable is a set of mutually exclusive states ordered or ranked in terms of degree of intensity that the states represent. Ordinals lack an unit of measurement but permits comparison of items which sometimes provide more important information required by social scientists. The variables can be graded as high, medium and low or in any other order form but they never tell how much high or low like interval data.

In economics, most of the consumer behaviour is analysed in terms of preferences: as high or low but cannot precisely indicate how much highly liked by the consumer for a commodity. Political scientists, sociologists and experimental psychologists very often use ordinal data for analysis where the opinion has to be analysed with regards to the behaviour of the persons. Sometimes the ordinal data is assigned a numerical value for comparison but it actually does not express the magnitude, it only states the nature of the ordinal data.

Inspite of difficulties of measurement, ordinal data is still regarded as important for many analyses. In many a case the concepts can be measured only in terms of ordinal data.

1. A consumer can express more of a particular set of items and less of other set which he feels in terms of order only. He cannot exactly say how much he feels for a commodity in comparison to other. But his preference is a truism.
2. In consumer preference analysis, the producer is not very much interested that how much the consumer spends on one item

or other, but which one he prefers to other so that accordingly the producer would determine the price of the commodity.

3. When the commodities are graded according to their quality as A, B, and C etc., it enables a consumer to select the commodity easily, but he never bothers what exactly is the difference between the two grades of the commodities. The grading also helps the producer to sell the commodities easily.
4. A student's intellects are graded as A, B, and C etc., and one knows that A grade is better than B grade, they are not equal, but what exactly is the difference between the two grades one cannot say.
5. In some situations only the ordering of observations on a qualitative variables matters, specific numerical value has no importance. In a football match it is not the score of goals that matters, but the position of the team on the overall game that is important.
6. A politician is not evaluated according to his age but according to his work performance and social popularity in the nation. Here again one cannot exactly say what are the qualitative difference between the two politicians.

Ordinal Association

Estimating an association measure poses specific problems for the ordinal variables. The main questions are:

1. How to find out that the ordinal variables are related?
2. If they are related how to find out a measure of association?
3. Whether one ordinal variable is useful in predicting the other?

New approaches are being used in order to answer the above questions by the investigators. An analysis of ordinal variables is presented here and explained with the help of cellular or box diagram.

Cross classification of two sets of ordinal variables comprising of 72 consumers are presented in Table 5.1. In X axis consumers are categorised according to their income range as higher, middle and lower income groups in the descending order and assumed as independent variables. There are 6 categories of dependent variables in the ascending order starting with vegetable pumpkin through tomato. According to the sample data collected in the cells the consumption pattern of each income group can be predicted.

There are two types of cells in the table—white and shaded. The white cells are success for prediction as they are the maximum frequencies in the respective income groups. The shaded cells are the error events. Let the statement X...... Y mean, "X tends to be a sufficient condition for Y". Hence according to the Table 5.1 prediction P can be stated in terms of white cells:

P : High ... Beans or tomato
Mid ... Potato or Gajar.
Low ... Pumpkin or Brinjal.

Table 5.1: Comparison of Ordinal Variables

Y : Consumption of vegetables.	X: Income Group			
	High	Middle	Low	
Pumpkin	0	0	9	9
Brinjal	2	3	11	16
Potato	3	8	8	19
Gajar	5	5	2	12
Beans	8	0	0	8
Tomato	8	2	0	10
Total =	26	18	30	4

To calculate the prediction its proportionate reduction in error has to be calculated. The reduction occurs when we make prediction for each case with knowledge of its independent variable state in contrast to predicting without such information. That a comparison is made of the observed error with the error rate of a bench mark prediction that does not make use of the independent variable information. This means first two errors have to be compiled to find out the association measure—Observed error and expected error.

Observed Error: Sum total of the values in the shaded cells are known as observed error. The observed errors are

High : 0 + 2 + 3 + 5 = 10
Mid : 0 + 3 + 0 + 2 = 5
Low : 8 + 2 + 0 + 0 = 10
8 + 7 + 3 + 7 = 25

Expected Error: In the first column of the Table there are 26 rich individuals as a simple, of which success observed predictions are beans and tomatoes. The rest of the four shaded cells are in errors: from Pumpkin to Gajar. In the whole population there are total of 56 individuals who consume the vegetables of inferior variety; from Pumpkin to Gajar and their marginal total 9 + 16 + 19 + 12 = 56. This is the error event for the bench mark replications for the higher income groups. This shows that an error will occur with probability of 56/74; since there are 26 replications predictions, the expected value of the higher income group would be 26 (56/74) = 19.68. Calculation of expected values for the three independent variables, according to the above calculation is shown in Table 5.2.

Table 5.2: Calculation of Expected Values

	Rich	Mid	Low
No. of replication predictions	26	18	30
Total observed error	56	43	49
Expected errors	19.68	10.46	19.86 = 50

The overall expected error can be calculated directly as

$$\frac{1}{74}[26 (56) + 18 (43) + 30 (49)] = 50$$

The prediction success P_3 is defined as:

$$P_3 = \frac{(\text{Expected Error}) - (\text{Observed Error})}{(\text{Expected Error})}$$

$$= 1 - (\text{Observed Error}/\text{Expected Error})$$

Accordingly the expected error for the Table 5.1 is:

$$P_3 = 1 - (25/50)$$

$$= 1 - 0.5 = 0.5$$

Probability Measure

Probability of each cell can be calculated on the basis of total sample n. Accordingly

$$p_{ij} = p_{21} = 2/74 = .03$$

Similarly the marginal probabilities can be calculated denoting:

Marginal Probability of row i as $p_{i\cdot}$

Marginal Probability of column j as $p_{\cdot j}$.

The marginal probabilities of rows of columns can be calculated as under:

$$P_{i\cdot} = \sum_{j=1}^{J} P_{ij} \qquad P_{\cdot j} = \sum_{i=1}^{I} P_{ij}$$

Accordingly marginal probabilities of the second row of the Table 5.1 is calculated as:

$$\begin{array}{ccccccc} P_{21} & + & P_{22} & + & P_{23} & = & P_2 \\ 2/74 & & 3/74 & & 11/74 & = & 16/74 \\ .03 & & .04 & & .15 & = & 0.22 \end{array}$$

The association measure can be calculated on the basis of individual error measures. Each cell in the table can be assigned measures as w_{ij}.

When the cell is a success $w_{ij} = 0$

When cell is an error : $w_{ij} = 1$

On this basis the observed and expected errors are calculated in the following formulas respectively.

$$\text{Observed error} = \sum_{i=1}^{I} \sum_{j=1}^{J} w_{ij}\, p_{ij}$$

$$\text{Expected Error} = \sum_{i=1}^{I} \sum_{j \neq 1}^{J} w_{ij}\, p_i\, p_j$$

Observed Errors: According to the above formula all the white cells are taken as zeros, hence the sum total of all the shaded cell's probability is the observed error of the table:

$$(2/74) + (3/74) + (3/74) + (8/74) + (5/74) + (2/74) + (2/74)$$
$$= 25/74 = .338$$

Expected Error: This calculation is done on the basis of marginal values, in respect of shaded cells only, hence the addition is

$$\begin{array}{ccccc}
(9/74)\,(26/74) & + & (9/74)\,(18/74) & & — \\
(16/74)\,(26/74) & + & (16/74)\,(18/74) & & — \\
(19/74)\,(26/74) & + & — & & (19/74)\,(30/74) \\
(12/74)\,(26/74) & + & — & & 0(12/74)\,(30/74) \\
— & + & (8/74)\,(18/74) & + & (8/74)\,(30/74) \\
— & + & (10/74)\,(18/74) & + & (10/74)\,(30/74)
\end{array}$$

$$= 3700/5476 = .676$$

Hence $p = 1 - (.338/.676) = 0.5$

Statistical Independence: When $p = 0$, that is when the observed errors are equal to the expected errors the variables can be said to be statistically independent. This will happen when $p_{ij} = p_i\, p_j$ in all the cells of the table.

Cohen's Kappa

Jocob Cohen first developed and used the measure of judgement reliability as Kappa (Greek for k). The main assumption is that it must be a square table, that is $I = J$. In other words both the variables must have same number of categories. If it is not a square matrix, it can be condensed into a square matrix by reducing rows or columns or both. The value of k depends on the nature of condensing the categories. The main diagonal values are taken as the success prediction. The Table 5.1, which is in 3 × 6 dimension is reduced into 3 × 3 dimension and shown as Table 5.3 below. Since the main diagonals have to be shown as success predictions the rows condensed have been changed positions. The last two rows shown as first row of the Table 5.3 and the first two rows shown as the last row in the condensed table in order to keep the values of the success cells in tact in the new form which is a condition for the calculation of kappa.

Table 5.3: Condensed Table of Table 5.1

	High	Middle	Low	
Bean + Tomato	16	2	0	18
Potato + Gajar	8	13	10	31
Pumpkin + Brinjal	2	3	20	25
Total =	26	18	30	74

In this table 16, 13, and 20 are the main diagonals which are treated as success prediction. Kappa is calculated as

$$k = \frac{\sum_{i=1}^{I} p_{ii} - \sum_{i=1}^{I} p_{i}\cdot p_{j}}{1 - \sum_{i=1}^{I} p_{i}\cdot p_{\cdot j}}$$

$P_{ii} = (16 + 13 + 20)/74 = 49/74 = .662$

$p_{i}\cdot p_{j} = (18/74)(26/74) + (31/74)(28/74) + (25/74)(30/74)$
$= 2086/5476 = .380$

$$= \frac{.662 - .380}{1 - .380}$$

$$= \frac{.282}{.62} = .455$$

If after the condensation some of the non-diagonal cells remain white, i.e. remain as success predictions, according to the formula it should be ignored or taken as error cells. In Cohen's kappa calculation all the non-diagonals are taken as error cells.

Quadrant Measure

W.H. Kruskal familiarised one quadrant measure to find out the association between the ordinal variables. In this method the cross-classified table is divided or condensed into 2 × 2 quadrant table for the purpose of analysis. The two main assumptions in this method are:

1. The analysis is confined to four cells only, in which the main diagonals are the success predictions and the other diagonals are the errors.
2. The table is condensed in such a manner that the error and success are 50-50, in case of both rows and columns.

A hypothetical table of ordinal variables are presented in Table 5.4 and its quadrant in Table 5.5. The table is condensed into four parts.

Table 5.4: A Hypothetical Table of Ordinal Variables

	X_1	X_2	X_3	
Y_1	1	0	1	2
Y_2	1	0	2	3
Y_3	0	3	0	3
Y_4	0	0	1	1
Y_5	0	0	1	1
Total =	2	3	5	10

The four quadrants in the Table 5.5 are x, x' and y, y'. The x is a combined category that contains half of the observations that are in the highest category on the independent variable and x'

combines the lowest categories. Similarly also for the dependent variables y and y'. It can be noticed that the entire table is divided into two parts both for dependent and the independent variables.

Table 5.5: The Quadrant Table

	X $(X_1 + X_2)$	X′ (X_3)	
$(Y_1 + Y_2)$: Y	2	3	5
$(Y_3 + Y_4 + Y_5)$: Y′	3	2	5
Total =	5	5	10

The formula for the Quadrant Measure Q. M. is

$$\text{Q.M.} = 1 - \frac{\text{(Observed error)}}{\text{(Expected error)}}$$

$$\text{Observed error} = p_{yx}{}' + p_{y}{}'_{x}$$

$$\text{Expected error} = p_{y\cdot}p_{\cdot x}{}' + p_{x\cdot}p_{y}{}'.$$

Since the table is divided as 50-50 the expected error can be taken as: (1/2) (1/2) + (1/2) (1/2) = (1/2); hence

$$\text{Q.M.} = \frac{p_{yx}{}' + p_{y}{}'_{x}}{(1/2)}$$

$$= 2\,(p_{yx}{}' + p_{y}{}'_{x})$$

Since $p_{yx'} = p_{y'x}$ then $2\,(p_{yx}{}') + 2(p_{y}{}'_{x}) = 4(p_{yx}{}')$; therefore the Quadrant Measure can be simply stated as:

$$\text{Q. M.} = 1 - 4\,(p_{yx}{}')$$

According to this formula the Q.M. for the Table 5.5 is:

$$\text{Q.M.} = 1 - 4\,(3/10)$$

$$= 1 - 4\,(.3) = -0.2$$

The main difficulty of this measure is that it is split into 50-50 and analysed; but in reality this may not happen in many of the cases.

Modified Measure: When it is not possible to divide a table into 50-50 still it can be analysed with certain modifications. Under modified measure the table is divided into 3 × 3 condensed form and the four corners of the table is taken as quadrants for measure.

The Table 5.1 which is condensed into 3 × 3 table and shown as Table 5.3 can be taken for further analysis under quadrant measure in Table 5.6. In this table the second row and the second column are eliminated for the purpose of quadrant analysis. In the table the two diagonal cells i.e. y_1x_1 and y_3x_3 are the success predictions and the other two cells y_1x_3 and y_3x_1 are the error cells which are shaded. The dotted cells in the table are completely eliminated from the analysis.

Table 5.6: Quadrant Analysis

	X_1	X_2	X_3	
Y_1	16	2	0	18
Y_2	8	13	10	31
Y_3	2	3	20	25
Total =	26	18	30	74

The formula for the Modified Quadrant Measure (MQM) is as follows:

$$\text{MQM} = 1 - \frac{p_{13} + p_{31}}{p_{1\cdot}\,p_{\cdot 3} + p_{3\cdot}\,p_{\cdot 1}}$$

Accordingly the MQM for the above table is:

$$\text{MQM} = 1 - \frac{(0/74)\,(2/74)}{(18/74)\,(30/74) + (25/74)\,(26/74)}$$

$$= 1 - (.027)/(.218) = .876$$

A major weakness of the quadrant measure, even under the modified version is that it evaluates a single fixed prediction.

6

Factor Analysis

There are several methods in statistics to find out possible relationships between and among the variables. One such method is factor analysis. K. Pearson early in this century developed the system of Principal Component Analysis and later H. Hoteling widely used this method in psychology. It was Charles Spearman in 1904 who introduced factor analysis to measure general intelligence and further developed it in 1926 as a separate method of analysis. The main difference between the factor analysis and the principal component analysis is that the latter is based on variance while the former on co-variance or correlation. Besides this, the factor analysis involves a hypothesis of linearity.

What is Factor Analysis?

Factor analysis is synthesis of variables – it analyses distinct factors at work among the variables. These new entities are themselves variables, hypothetical variables, which are fewer than the raw variables. The purpose of the factor analysis is therefore to find out these common factors which provides the shape of real structure hidden in the multiplicity of the variables.

Factor analysis is a wholistic method of analysing multiple variates. But it is different from the partial and multiple correlation on two accounts:

1. Factor analysis can be compared to partial correlation. It is same as partial correlation except that in partial correlation the other variables are assumed constant while in factor analysis the other factors are assumed as constant.

2. Factor analysis is like a multiple correlation, but more systematic than the multiple correlation. In factor analysis, first variables are grouped to give estimate of independent functional units and then proceeds for analysis as in case of a multiple correlation.

Main Characteristics

Factor analysis attempts to determine the quantitative relationship between the variables where the relationships are due to certain general causal factors. The specific characteristics of the factor analysis are:

(i) Factor analysis is concerned with how much relationship exists, while other analyses indicate only any significant relationship that exists.

(ii) Factor analysis is a fine quantitative tool. It explains both how many are in action and what is the magnitude of their action.

(iii) Factor analysis yields more extensive evidence of other kinds of interaction, extent of assistance or opposition of influence in their effects.

This shows that a factor analysis "without hypothesis formation" can arrive at highly structured answers regarding:

(a) Number of factors at work,

(b) Nature of factors,

(c) Their correlation and

(d) Their magnitude to the variance of particular variables or most of the variables.

Scientific Research

Four important enquiries are involved in a scientific research and in factor analysis all the four enquiries are taken up with fewer assumptions. The four enquiries are:

1. In scientific research controls are operated on experimental basis or the behaviour is observed for analysis.

2. A researcher chooses a collection of variables arbitrarily and their relationship is investigated or their importance established by their empirical correlation.

3. A researcher requires a hypothesis for tests and investigations and sometimes he finds out or produces a hypothesis out of observation.

4. Sometimes the scientific analysis forces the researcher to make prior assumptions with regard to the independent and the dependent variables; on other occasions it leaves the causal relationship unassumed and open for later proves. All these four aspects of scientific enquiry is found in factor analysis and hence it is regarded as one of the important analysis in both physical and social sciences.

Aims of Factor Analysis

First, factor analysis helps to obtain a smaller sample from the original materials. Instead of observing large number of variables, fewer factors which are common to the variables are analysed, which makes the problem easier and clearer.

Second, when common factors are extracted from a set of variables it gives a separate pattern of motion between variables for analysis. This cannot be found in the analysis of the raw variables.

Third, instead of large number of variables the pattern of motion of the factors can be interpreted logically for all types of relationships of the variables.

Hence the factor analysis is an important aid in observing the mutual relationships between the variables. These relationships can be formulated as a theory or theories can be tested on factor analysis. Relationships can be explained in a regression analysis but in factor analysis the advantage is that it can indicate the number of equations that are required for a solution. Another advantage is that it helps comparative analysis. When various index numbers are prepared on the basis of common factors the comparison is possible and easier too.

Types of Factor Analysis

According to the purpose of analysis the factor analysis can be classified into two: (1) Exploratory Factor Analysis and (2) Confirmatory Factor Analysis. .

Exploratory Analysis: In this analysis the researcher starts analysis without knowing the number of underlying common factors that are present in the observed data. The researcher tries to extract minimum number of factors that explains the major proportions of the observed variables. Hence this is known as exploratory analysis. This method is mostly used in economics and other social sciences.

Confirmatory Analysis: This method is used primarily for testing the hypothesis. If an investigator feel that there are two common factors which influences the behaviour of the consumers among the number of variables that explains the behaviour; this can be taken as hypothesis for testing. If factor analysis is used to confirm such factors among the variables it is known as confirmatory factor analysis. Even though the two methods are separate, one may use both the methods to evaluate a particular problem.

Factor analysis is also used as a basis for further investigations. Before conducting other analysis one may first perform the factor analysis to find out the number of variables that are effective for analysis. A factor analysis reveals the number of variables that are not influential on the underlying factors and accordingly a researcher can eliminate these variables altogether from the investigation.

Characteristics of the Variables

Factor analysis is fundamentally based on the assumption that there are certain underlying common factors which influence the variables. The number of factors which influence the variables are generally less than the number of variables. If the number factors are same as the variables or more there would be no necessity of factor analysis because it is easier to analyse the observed variables rather than the factors, since the former are less in number. The purpose of the factor analysis is to find out the common factors in the variables and select some of the most influential factors for analysis, in order to know the behaviour of the variables.

The first task of a factor analyst is to find out what and how the variables are composed of the common factors. For this a variable can be divided into three components: (1) Common factors (2) Unique factors and (c) Disturbance term.

A set of variables taken for analysis, on the basis of random sampling or in other methods, such as X_1, X_2, X_3, X_4, etc., are known as observed variables. These variables are dissected for the purpose of extraction of common factors.

1. ***Common Factors:*** It is assumed that the observed variables possess certain common characteristics, common to two or more observed variables, called common factors or simply factors. One observed variable may be composed of one or more factors in common with other variables.

Table 6.1: Composition of the Observed Variables

$$X_1 = F_1 + 0 + U_1 + e_1$$

$$X_2 = F_1 + 0 + U_2 + e_2$$

$$X_3 = F_1 + F_2 + U_3 + e_3$$

$$X_4 = F_1 + F_2 + U_4 + e_4$$

$$X_5 = 0 + F_2 + U_5 + e_5$$

$$X_6 = 0 + F_2 + U_6 + e_6$$

Table 6.1 shows two common factors F_1 and F_2 that explains the behaviour of the observed variables X_1 to X_6. The F_1 in common to the variables X_1 to X_4 and F_2 is common to the variables X_3 to X_6 and both the factors are common to two variables, that is, X_3 and X_4. The factors are also variables. They are called 'source variables'. Unlike the observed variables the factors are hypothetical variables.

The relationship of the observed variables and the source variables presented in Table 6.1 is shown again in the form of a path diagram in Figure 6.1. The solid lines indicate the influence of the factor concerned to the observed variables. The error terms are not shown in the figure because it is assumed that the collection of the data is perfect.

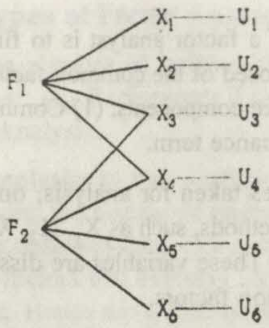


Fig. 6.1 Path Diagram for Orthogonal Model

Addition of the factors with their coefficients are called communality and it is denoted as h^2. For variable X_3 the communality is:

$$X_3 = (a_1F_1 + a_2F_2) = h_3^2$$

in which a_1 and a_2 are the coefficients of F_1 and F_2 respectively. The factor coefficients and communalities are the main basis for factor analysis.

2. ***Unique Factors:*** The second component of the observed variable is called unique factor: U. This factor is exclusive to only one variable; this character is not found in any of the other variables. Hence every observed variable is divided into two parts, the factors or characteristics common to all or some of the variables in the set; and the other part exclusive to a particular variable or uniqueness which is indicated as U.

3. ***Error Term:*** The third component of an observed variable is the error term. In a statistical analysis the unexplained part is always provided as disturbance term or error term. This error term is denoted as e. If one collects error free data or there is no part that remains unexplained beyond common and the unique factors the error term can be eliminated from the analysis. In this discussion, the error term has not been utilised with the assumption that there are no errors in the observed variables.

Assumptions

Linearity: Factor analysis assumes that the observed variables are linear combinations of underlying factor or hypothetical variables.

Hence every observed variables is a linear combination of common and unique factors. Only the common factors are known as factors and it is their analysis that is known as factor analysis. Hence the unique factors are not taken into account in the factor analysis, and so every ovserved variable is influenced by the combination of common factors. This is shown in Table 6.2 in the matrix notation. Here 'a's are the constant coefficients of the factors.

Table 6.2: Factor Combinations in Matrix Form

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} a_{11}\ a_{12}\ a_{13} \\ a_{21}\ a_{22}\ a_{23} \\ a_{31}\ a_{32}\ a_{33} \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ F_3 \end{bmatrix}$$

Relationship of Unique Factors: In the factor analysis the main aim is to analyse the common factors hence the unique factors are eliminated from the analysis. In other words, it is assumed that the unique factors have no relationship with the common factors, that is, there is no relationship between F_1, F_2, F_3, and U, and also between U_1 and U_2 or U_1 and U_3 or U_2 and U_3. Mathematically the covariance between F's and U's and U's and U's are all zero:

$$\text{cov.}\ (F_1U_1) = \text{cov.}\ (F_2U_2) = \text{cov.}\ (F_3U_3) = 0$$

$$\text{cov.}\ (U_1U_2) = \text{cov.}\ (U_1U_3) = \text{cov.}\ (U_2U_3) = 0$$

Standardisation of Variables: All the observed variables, $X_1X_2X_3$; source variables or common factors $F_1F_2F_3$ and unique factors U_1, U_2, and U_3 are assumed as standard variables. Hence their mean is zero, and standard deviation and variance is equal to 1. Accordingly the total variance of a standard variable X_1 is divided into two parts: (1) Communality h^2 and (2) Uniqueness U^2 and therefore

$$X_a{}^2 = h^2 + U^2$$

This shows that the total variance of the observed variable is equal to the sum total of the variance of common factors and the unique factors. As explained earlier the communality h^2 is the total of the squares of the coefficients of the factors. That is

$$a_1{}^2 F_1 + a_2{}^2 F_2 = h^2$$

Covariance and Correlation

Covariance and correlation are the two main basis for the factor analysis. It is known that mean and variance are the two important properties of variables. The mean indicates the value of the central tendency while the variance indicates the degree of dispersion of the variables. Variance between two variables is known as covariance. It measures the extent to which the value of one variable tends to covary with the value of other variable.

For the simplicity of the equations the expected notation E is used here as an abbreviation for "adding all the values and dividing that sum by the total number of cases.

Instead of $\overline{X} = \Sigma X / N$, it is simply noted as $\overline{X} = E(X)$

$$\text{Variance of } X_1 = \frac{\Sigma(X_i - \overline{X})^2}{N} \text{ or } E\,(X - E(X))^2 \qquad (1)$$

$$\text{Co-variance of } (F_1 X_1) = \Sigma\left\{(F_1 - \overline{F}_1)(X_1 - \overline{X}_1)\right\} / N$$
$$= E\ \{[F - E(F)]\ (X = E(X))\} \qquad (2)$$

When the variables are standardised the mean are assumed as zeros. In the equations (1) and (2) when $X_1 = 0$ and $F_1 = 0$, the equations can be written as:

$$\text{Var. } (X_1) = E(X_1)2 \qquad \ldots (3)$$

$$\text{Cov. } (F_1 X_1) = E(F_1 X_1) \qquad \ldots (4)$$

F_1 is the factor and the X_1 is the observed variable in the set of variables taken for factor analysis. As it is known that $X = F + U$ hence

$$\text{Var. } X = V_x = E\ (aF + bU)^2$$

$= E\ (a^2F^2 + b^2U^2 + 2abFU)$ Since expectation of constant is a constant hence

$$= a^2E(F^2) + b^2E(U^2) + 2ab\ E(FU)$$

Since Cov. (FU) = 0 as there is no correlation the equation is reduced to:

$$= a^2 \text{ var. } F + b^2 \text{ var. } U$$

This shows that variance of X_1 is summation of the following two components:

1. The variance of F times the square of the coefficient associated with it.
2. The variance of U times the square of the coefficients associated with U.

Since the observed variables are standard variables its variance is equal to 1. Hence:

$$\text{Cov. }(X_1) = a^2 + b^2 = 1 \quad \ldots (5)$$

Nature of Co-variance

$$\text{Cov. }(F_1X_1) = E\ (F_1X_1) \quad \ldots(6)$$

$$\text{Since } X_1 = (a_1F_1 + b_1U_1)$$

$$\text{Cov. }(F_1x_1) = E\ \{(F_1)\ (a_1F_1 + b_1U_1)\}$$
$$= a_1E(F_2) + b_1E(F_1U_1) \quad \ldots (7)$$

Since there is no relationship between F and U, hence the covariance between them is zero, so the above equation (7) can be finally written as

$$\text{Covariance }(F_1X_1) = a_1 \text{ var. } F_1$$

Since the observed variables are all assumed as standardised variables the co-variance is called the correlation coefficient; r; and so its magnitude varies between –1 to +1. Accordingly the coefficient of determination R^2 is always remain within 1. Covariance between F_1 and X_1 is then the correlation between the two, that means:

$$\text{Cov. }(F_1X_1) = rF_1X_1 = a_1$$

In this sense a_1 therefore can be accepted as multiple regression coefficient X_1 on F_1 and U_1. The linear weights 'a' and 'b' are the standardised regression coefficients. The X is here a dependent variable and Fs and Us are the independent variables.

Accordingly correlation between any two observed variables say between X_1 and X_2 can be stated as covariance of X_1 X_2

$\text{Cov.}\ (X_1X_2) = E\ (X_1X_2)$
$X_1 = (a_1F_1+b_1U_1)$ and $X_2 = (a_2F_1 + b_2U_2)$
$\text{Cov.}\ (X_1X_2) = E\ (a_1F_1 + b_1U_1)\ (a_2F_1 + b_2U_2)$
$= a_1a_2\ X\ \text{var.}\ F_1$

Types of Analysis

On the basis of relationship among the factors, the factor analysis can be divided into two types: (1) Orthogonal and (2) Oblique.

1. Orthogonal Model

In the orthogonal factor model it is assumed that there is no correlation between the factors in the analysis. In a two factor model explained above it is assumed that there is no correlation between F_1 and F_2 is zero. In other words the covariance of F_1F_2 is zero.

$\text{Cov.}\ (F_1, F_2) = 0$

In geometrical analysis the axes X and Y forming 90° angle represents no correlation. The factor model presented in Table 6.1, and shown in path diagram in Fig. 6.1 is an orthogonal model, since there is no correlation between F_1 and F_2.

2. Oblique Model

In oblique factor model where more than one factor is involved, correlation between the factors are also taken into consideration. In a three factor model the following factor correlation matrix is taken for analysis, as shown in Table 6.3.

Table 6.3: **Factor Correlation Matrix**

	F_1	F_2	F_3
F_1	1	.6	.4
F_2	.6	1	.3
F_3	.4	.3	1

Figure 6.2 depicts a path diagram for an oblique factor model as per the correlation of coefficients presented in the correlation matrix. In the oblique model the correlation can be shown in a X. Y plane

in a straight line less than 90° angle. The correlation is cosign theta, (cos. θ) of a particular angle.

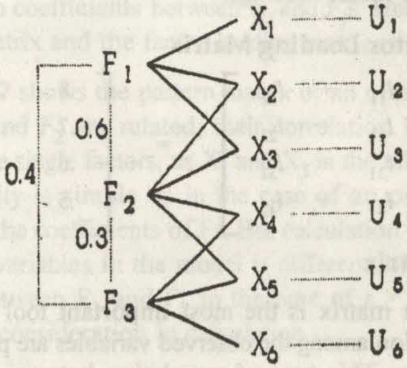


Figure 6.2 Path Diagram for Oblique Model

Orthogonal Model

Factor Loadings

When both sets of variables, the observed variables; X_1, X_2, X_3, etc., the hypothetical variables, F_1, F_2, F_3, etc. and the Us are all standardised and have unit covariance the linear coefficients of the factors a_1, a_2, a_3, are called factor loadings. In a regression analysis the same coefficients are known as standard regression coefficients. Since the total variance of $X_1 = 1$, so $h^2 + U = 1$. As we knew earlier $(F_1)^2 + (F_1)^2 = h^2$. Table 6.4 shows the factor loadings and the communality together with Us coefficients which are equal to one. For instance in row one $h^2 + U = .89 + .11 = 1.00$.

Table 6.4: Communality and Factor Loadings

Observed variables	*Factor F_1*	*Loadings F_2*	*Communality h^2*	*Uniqueness U*
X_1	.8	.5	.89	.11
X_2	.7	.4	.65	.35
X_3	.6	.4	.52	.48
X_4	.5	.5	.50	.50

The columns of F_1, F_2, in table 6.4 is known as structure Matrix or Factor Loading Matrix. In algebraic form it can be stated as under, where *'l'* denotes factor loadings.

Table 6.5: Factor Loading Matrix

$$F = \begin{bmatrix} l_{11} & + & l_{12} \\ l_{21} & + & l_{22} \\ l_{31} & + & l_{32} \\ l_{41} & + & l_{42} \end{bmatrix} = \begin{bmatrix} .8 & .5 \\ .7 & .4 \\ .6 & .4 \\ .5 & .5 \end{bmatrix}$$

Correlation Matrix

Correlation matrix is the most important tool for the factor analysis. Correlation among the observed variables are placed in matrix form for analysis. This type of correlation between X_1 and X_2 is not the usual statistical correlation, where one variable is taken as direct cause of other variable. Here correlation between the two observed variables are analysed in terms of their common factors. Hence the correlation between X_1 and X_2 is computed on the basis of factor loadings of the respective variables. Correlation of $X_1X_2 = r_{12} = l_{11}l_{12}$ (For one factor model)

$= l_{11}l_{21} + l_{12}l_{22}$ (For one factor model)

On the basis of the factor loadings of orthogonal model presented in Table 6.5, the correlation among the observed variables are computed in the following manner and presented in Table 6.6.

$$r_{12} = r_{21} = (l_{11}l_{21}) + (l_{12}l_{22}) = (.8)(.7) + (.5)(.4) = .76$$
$$r_{13} = r_{31} = (l_{11}l_{31}) + (l_{12}l_{32}) = (.8)(.6) + (.5)(.4) = .68$$
$$r_{14} = r_{41} = (l_{11}l_{41}) + (l_{12}l_{42}) = (.8)(.5) + (.5)(.5) = .65$$
$$r_{24} = r_{42} = (l_{21}l_{41}) + (l_{22}l_{42}) = (.7)(.5) + (.4)(.5) = .55$$
$$r_{23} = r_{32} = (l_{21}l_{31}) + (l_{22}l_{32}) = (.7)(.6) + (.4)(.4) = .58$$
$$r_{34} = r_{43} = (l_{31}l_{41}) + (l_{33}l_{42}) = (.6)(.5) + (.4)(.5) = .50$$

Table 6.6: Correlation Matrix of Factor Loadings

	X_1	X_2	X_3	X_4
X_1	1	.76	.68	.65
X_2	.76	1	.58	.55
X_3	.68	.58	1	.50
X_4	.65	.55	.50	1

Oblique Model

In the oblique model correlation between the two or more factors are taken into consideration in computing communality as well as the correlation coefficients between X_i and F_is. Hence in this model the pattern matrix and the factor loading matrices are different.

Table 6.7 shows the pattern matrix of an oblique factor model in which F_1 and F_2 are related; their correlation being $rF_1F_2 = .6$. When there are single factors, as X_1 and X_5 in the table the calculation of communality is simple as in the case of an orthogonal model; i.e. square of the coefficients of Fs. But calculation of communalities for the other variables in the model is different since there is inter-correlation between F_1 and F_2 to the tune of $r = .6$; which has to be taken into consideration in calculation.

Table 6.7: Pattern Matrix and Communality

Variables	F_1	F_2	h^2	U
X_1	.8	—	.64	.36
X_2	.5	.4	.65	.35
X_3	.3	.2	.21	.79
X_4	.3	.5	.52	.48
X_5	—	.6	.36	.64

There is a separate formula to calculate the communality in case of oblique model:

$$h^2 = F_1^2 + F_2^2 (F_1 F_2)(rF_1 F_2)(2)$$

Accordingly the communality of X_2 is calculated as:

$$X_2 = h^2 = (.5)^2 + (.4)^2 + (.5)(.4)(.6)(2)$$
$$= (.25) + (.16) + (.24)$$
$$= 0.65$$

Structure Matrix

In the oblique model the factor loading matrix is called the structure matrix. In case of orthogonal model the pattern matrix is also the factor loading matrix. But in case of oblique model the factor loading matrix has to be calculated afresh, known as structure matrix (in order to avoid confusion of factor loading matrix of orthogonal model) on the basis of pattern matrix reflecting relationship between the factors F_1 and F_2. Hence rF_1X_1 and rF_2X_1 have to be calculated again taking $rF_1F_2 = .6$ into consideration. The formula for the calculation of factor loadings are:

$$rF_1X_1 = l_{11} + l_{12}\, rF_1F_2$$
$$rF_2X_1 = l_{11}\, rF_1F_2 + l_{12}$$

On the basis of the above formula F_1 and F_2 are calculated in Table 6.8 (a) and (b) respectively and the results are presented in the structure matrix in Table 6.9.

Table 6.8(a): Calculation of F_1

$rF_1X_1 = (.8) + \text{——} = .80$

$rF_1X_2 = (.5) + (.4)(.6) = .74$

$rF_1X_3 = (.3) + (.2)(.6) = .42$

$rF_1X_4 = (.3) + (.5)(.6) = .33$

$rF_1X_5 = \text{—}\ (.6)(.6) = .36$

Table 6.8(b): Calculation of F_2

$rF_2X_1 = (.8)(.6) + \text{—} = .48$

$rF_2X_2 = (.5)(.6) + (.4) = .70$

$rF_2X_3 = (.3)(.6) + (.2) = .38$

$rF_2X_4 = (.3)(.6) + (.5) = .68$

$rF_2X_5 = \text{——}\ (.6) = .60$

Fresh calculation of factor loadings on the basis of pattern matrix above are presented in a matrix form below as structure matrix. It can be seen that all the elements of the structure matrix is filled up, while in the pattern matrix even though there was no correlation between X_1F_2 and between X_5F_1 where the two cells (Table 6.7) remained empty. Further it can be found that in comparison to the pattern matrix all the elements in the structure matrix have higher values.

Table 6.9: Structure Matrix

	F_1	F_2
X_1	.80	.48
X_2	.74	.70
X_3	.42	.38
X_4	.33	.68
X_5	.36	.60

Correlation Matrix

Next step in oblique model analysis is calculation of correlation matrix of observed variables in terms of structure matrix. The formula for calculation of correlation coefficient is same as in the case of orthogonal model. As stated above the formula is

$$
\begin{aligned}
\text{Correlation } X_1X_2 = r_{12} &= (l_{11}l_{21}) + (l_{12}l_{22}) \\
&= (.80)\,(.74) + (.48)\,(.70) \\
&= 0.59 + 0.34 \\
&= 0.93
\end{aligned}
$$

Accordingly the correlation matrix of the oblique model is exhibited in Table 6.10.

Table 6.10: Correlation Matrix

	X_1	X_2	X_3	X_4	X_5
X_1	—	.93	.52	.59	.58
X_2	.93	—	.58	.72	.69
X_3	.52	.58	—	.40	.38
X_4	.59	.72	.40	—	.53
X_5	.58	.69	.38	.53	—

Extraction of Factors

After preparation of the correlation matrix the next phase of the factor analysis is the extraction of factors. There are several methods to extract factors; chief among them are (1) Maximum likelihood method, (2) Least Squares method, (3) Alpha Factoring, (4) Image analysis and (5) Principal Component method.

A simple system of diagonal extraction of factors based on principal component method is discussed here for practical extraction of factors.

Adjusted Correlation Matrix

When a correlation matrix of the observed variables in terms of the factor loadings is constructed, it is ready for the extraction of the factors. In the correlation matrix, when the diagonal cells are replaced by the communalities of the respective variables the matrix is known as "Adjusted Correlation Matrix".

For the larger matrices the communalities will not make any change in the operation, but for the smaller ones, generally less than 12 variables, the communalities have to be inserted for an adjusted correlation matrix.

As shown earlier, the communalities can be calculated from the pattern matrix. There are easier methods of computing communalities also. Sometimes when the correlation matrix is not in monotonic form the communalities may pose problem in computation of factor loadings for the extraction of factors from the adjusted correlation matrix. One easiest method is to replace the diagonals with the highest value of each column as communality; without taking into account the signs of the values, positive or negative. Several arbitrary methods are also used by the researchers.

An example of Adjusted Correlation Matrix with the diagonals replaced with the communalities is presented in Table 6.11.

Table 6.11: Adjusted Correlation Matrix

	X_1	X_2	X_3	X_4
X_1	(.36)	.42	.48	.54
X_2	.42	(.65)	.68	.71
X_3	.48	.68	(.73)	.78
X_4	.54	.71	.78	(.85)

The total relationship of observed variables with the factors, both common and unique factors, can be denoted as R. Since the unique factors are not taken for analysis, the Adjusted Correlation Matrix is denoted as R–U^2 matrix.

The F_1 vector is the factor loading of the first extracted factor. The procedure of extraction is $(R-U^2) - F_1 F'_1$. Multiplication of column vector F_1 with its transpose row vector F'_1 gives the $F_1F'_1$ matrix as shown in Table 6.12.

Table 6.12: Calculation of F_1F_1' Matrix

$$\begin{bmatrix} .60 \\ .70 \\ .80 \\ .90 \end{bmatrix} [60 \quad .70 \quad .80 \quad .90] = \begin{bmatrix} .36 & .42 & .48 & .54 \\ .42 & .49 & .56 & .63 \\ .48 & .56 & .64 & .72 \\ .54 & .63 & .72 & .81 \end{bmatrix}$$

When the $F_1F'_1$ matrix is subtracted from the matrix of $(R - U^2)$ the residual matrix is R_1 $(R - U^2) - F_1F'_1 = R$. This is shown in Table 6.13.

Table 6.13: Extraction of First Factor and Residual Matrix

$$(R-U^2) - F_1F'_1 = R_1$$

$$\begin{array}{cc} & \begin{array}{cccc} X_1 & X_2 & X_3 & X4 \end{array} \\ \begin{array}{c} X_1 \\ X_2 \\ X_3 \\ X_4 \end{array} & \left[\begin{array}{cccc} .36 & .42 & .48 & .54 \\ .42 & .65 & .68 & .71 \\ .48 & .68 & .73 & .78 \\ .54 & .71 & .78 & .85 \end{array}\right] \end{array} - \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & X_3 & X_4 \end{array} \\ \left[\begin{array}{cccc} .36 & .42 & .48 & .54 \\ .42 & .49 & .56 & .63 \\ .48 & .56 & .64 & .72 \\ .54 & .63 & .72 & .81 \end{array}\right] \end{array}$$

$$= \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & X_3 & X_4 \end{array} \\ \left[\begin{array}{cccc} 0 & 0 & 0 & 0 \\ 0 & .16 & .12 & .08 \\ 0 & .12 & .09 & .06 \\ 0 & .08 & .06 & .04 \end{array}\right] \end{array}$$

After the extraction of the first factor the remaining elements in R_1 is called reduced Adjusted Correlation Matrix. It can be seen that in matrix R_1 the first column and the first row becomes zero, indicating that F_1 fully explains the behaviour of the observed variables in terms of the first factor. Instead of zeros if there are elements in the first row and column in some smaller values after extraction it indicates that the first factor is not fully explaining the behaviour of observed variables.

Extraction of the Second Factor: F_2

For the extraction of the second factor the same procedure as in case of first factor, is repeated. A column is selected which has highest sum of correlations or the second column for the extraction of the second factor. In R_1 matrix the second column is selected for extraction of F_2 which is incidentally has the highest sum of correlations.

In column 2, the diagonal value is .16 and its root over is .40. As a ratio to this value the factor loadings are calculated as under:

	Procedure		F_2
X_1	0	=	0
X_2	$\sqrt{.16}$	=	.40
X_3	.12/.4	=	.30
X_4	.08/.4	=	.20

On the basis of factor loading matrix $F_2F'_2$ is calculated in the following Table 6.14.

Table 6.14: The $F_2F'_2$ Matrix

$$\begin{bmatrix} 0 \\ .40 \\ .30 \\ .20 \end{bmatrix} [0 \quad .40 \quad .30 \quad .20] = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & .16 & .12 & .08 \\ 0 & .12 & .09 & .06 \\ 0 & .08 & .06 & .04 \end{bmatrix}$$

In the second extraction, $F_2F'_2$ is subtracted from R.$R_1 - F_2F'_2 = R_2$. The R_2 is again called the reduced Adjusted Correlation Matrix. Table 6.15 shows the second extraction.

Table 6.15: Extraction of the Second Factor

$$\begin{array}{ccccc} R_1 & & F_2F'_2 & & R_2 \\ \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & .16 & .12 & .08 \\ 0 & .12 & .09 & .06 \\ 0 & .08 & .06 & .04 \end{bmatrix} & - & \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & .16 & .12 & .08 \\ 0 & .12 & .09 & .06 \\ 0 & .08 & .06 & .04 \end{bmatrix} & = & \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

As it can be seen that after the extraction of the two factors the adjusted correlation matrix becomes a 0 matrix. This shows that the four observed variables are fully explained by the two factors. If after the extraction of the two factors still some of the elements remain positive further extraction of the factors may be taken. In practical cases after the extraction of the factors some smaller values are still found in the adjusted correlation matrix which are generally ignored as insignificant.

Degree of Explanation

The degree of influence of F_1 and F_2 can be calculated on the basis of the communality computed from the extracted factor loadings. This is shown in Table 6.16.

Table 6.16: Calculation of Degree of Influence

F_1	F_2	F_1^2	F_2^2	h^2
.60	0	.36	0	.36
.70	.40	.49	.16	.65
.80	.30	.64	.09	.73
.90	.20	81	.04	.85
	Total	2.30	29	= 2.59

The two factors together explain 2.59/4 = .65 per cent. Individually F_1 explains 2.3/4 = 0.58 or 58 per cent.

F_2 explains 0.29/4 = 0.07 or 7 per cent.

Of course of the total explanation of 2.59 each factor contributes F_1 = 89 per cent and F_2 = 11 per cent.

For the correlation matrix given in Table 6.10, after the extraction of two factors, there remains still a residual matrix as shown in Table 6.17. Some elements remained unexplained which are not concerned to the factors.

The factor loadings of the two extracted factors from the above matrix together with their communalities:

Table 6.17: The Residual Matrix after the Extraction of Two Factors

(a) From Table 6.10

	X_1	X_2	X_3	X_4	X_5
X_1	.01	0	0	0	0
X_2	0	–.05	.04	.08	0
X_3	0	.04	.28	.04	0
X_4	0	.08	.04	.18	0
X_5	0	0	0	.10	.01

(b)

F_1	F_2	$F^2{}_1$	$F^2{}_2$	h^2
.95	0	.90	0	.90
.98	.19	.96	.04	1.0
.55	.11	.30	.01	.31
.62	.28	.38	.08	.46
.61	.58	37	.34	.71
	Total	2.91	.47	3.38

In the above extraction of factors both the factors together explain 3.38/5 = .68 or 68 per cent; and each factor individually F_1 = 58 per cent, and F_2 = 9 per cent.

In an exploratory factor analysis, since a researcher does not know about the extract number of factors underlying in the observed variables, one has to proceed further extraction of the factors until all the elements of the adjusted correlation matrix becomes zero or negligible.

Retention of Factors

Retention of number of factors is very important for a researcher. After the extraction of the factors the researcher has to select the number of factors to be retained for analysis. If there are factors which explain only small fraction of the observed variables it is better to eliminate them from the study. Most of the factor analysts use eigen values as the index of retaining factors. The sum of the squared factor loadings of the extracted factors are called eigen values. The column totals of F_1 and F_2 in Table 6.16 and Table 6.17 (b) are the eigen values.

The General Principle of Retention

1. The factors are retained for analysis which have the eigen values greater than 1. This is in case of correlation matrix not adjusted. This works well in population correlation matrix. The number of factors responsible for the correlation matrix will always be equal to or greater than the number specified by this criterion.
2. Vectors in which eigen values greater than zero is retained when the reduced correlation matrix (with squared multiple r's in the main diagonals) is decomposed. But this is not applicable for the sample correlation matrix.
3. When the eigen values are negative, the above procedure is not applicable because to retain all the vectors more than zero becomes meaningless. Herman has suggested another principle in this case: stop extraction of factors before the cumulative sum of eigen values exceeds the sum of estimated communalities.
4. On the basis of the propcrtion of the eigen values to the total variance of the analysis one can determine the number of factors to be retained. Of course one has to decide at what proportion the factors have to be retained, i.e. at 1, 5 or 10 per cent of the total value.

$$\text{Proportion} = \frac{\text{Sum of the main diagonals}}{\text{Squares of factor loadings}}$$

Scree Test

Scree test is a geographical method familiarised by R.B. Cattell in biometrics. A graph is drawn between the number of factors and eigen values as shown in Figure 6.3. The principle is to stop factoring at the point where the eigen values begin to level off forming a straight line with an almost horizontal slope. This smooth slope is described as "factorial litter" or scree. Scree is a geographical term referring to a debris which collects on a lower part of a rocky slope. Many accept this method as good approach, but it is a subjective method.

In Fig. 6.3 eigen values are measured in the Y axis and the factors extracted from the data set in the X axis. Accordingly a curve is drawn. At point F the curve takes a clear turn and form there it is comparatively a smooth one, which runs almost along with the X axis. The F point indicates factor 4, hence only up to four factors have to be retained and rest should be eliminated from the analysis.

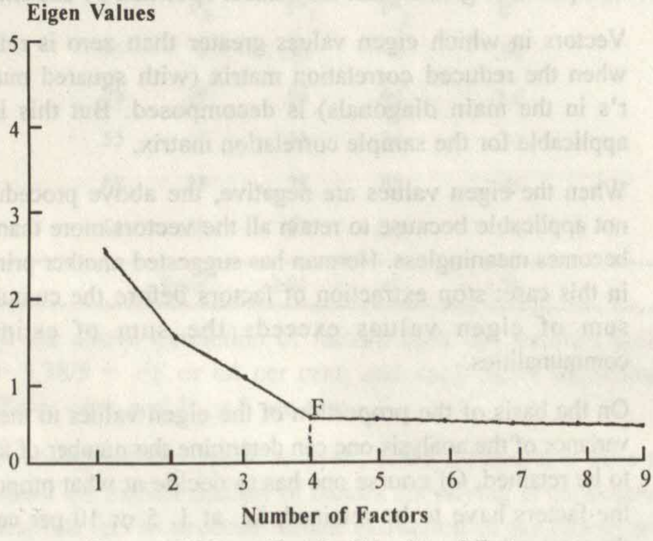


Figure: 6.3 Scree Test for Selection of Factors

The Method of Least Squares

There are two least square methods used in factor analysis. One is the principal axis factoring with iterated communalities and second is the method of minors developed by G.S. Thompson and H.H. Herman respectively.

General procedure for obtaining solution in least square approach are:

1. Assume k factors that account for the observed correlation.
2. Make initial estimates of communalities. This is the squared multiple correlation between a variable and remaining variables which is used as communality.
3. Extract k factors that can best reproduce the observed correlation matrix. On the basis of the methods of least squares, the equation to be solved is det (R-AI) = 0.
4. In order to get the factors patterns the communalities are reestimated on the basis of factor patterns obtained in the previous stage.
5. The process is repeated until no improvement is made in the factor structure.

There is another method of finding out the number of factors underlying in a data set. This is based on maximum likelihood analysis which assumes that the variables have a multivariate normal distribution.

First this equation is estimated $\hat{F}\hat{F}' + \hat{U}^2 = \hat{R}$. Where $\hat{R}$ is the estimated correlation matrix based on the maximum likelihood estimates of the factor loading matrix. R is the single correlation matrix. Fit one factor to the correlation matrix and test the goodness of fit with chi square χ^2 approximation. Closer the $\hat{R}$ to R better is the fit.

If χ^2 is not significant then that hypothesis of one factor is accepted. If significant (indicating lack of fit) two factors are fitted to the correlation matrix and goodness of fit for two factor model is tested by using the χ^2 test. If it is insignificant then two factor model is accepted. Otherwise a three factor solution is fitted and tested; this procedure continues until χ^2 is significant. In this manner the number of factors that are required for analysis is determined.

Number of Underlying Factors

In an exploratory factor analysis the researcher does not know about the number of underlying common factors in a data set. But this can be known with the help of the rank theorem of matrices. The adjusted correlation matrix has certain fundamental structural characteristics which is known as rank of a matrix. The degree of linear dependence in a set of vectors forming the matrix is referred as the rank of a matrix. The rank indicates the number of underlying factors in the data set.

Two simple principles of the rank theorem is explained here to understand the concept.

1. Rank of a matrix one, means all the determinants involving two or more variables should be zeros. The determinant involving two variables means:

   $$\det \begin{bmatrix} a^2{}_1 & r_{12} \\ r_{21} & a_2{}^2 \end{bmatrix} = a_1{}^2 a_2{}^2 - r_{12} r_{21} = 0$$

   When the rank of a matrix is one then the number of common factors involving the data is also one. In other words one common factor explains the observed variables.

2. When the rank of a matrix is two then there are two factors that explain the behaviour of the observed variables. When at least one determinant in the matrix involving two columns and rows will not be zeros, but more than two variables become zeros it is known as rank of the matrix. Like that when the rank of a matrix is 3 there are three common factors and so on. In other words if rank of a matrix k means the underlying factors are also k.

On the basis of rank theorem the communalities of an observed variable can be calculated.

$$\text{Since } a^2{}_1 a^2{}_2 - r_{12} r_{21} = 0$$

$$a^2{}_1 a^2{}_2 = r_{12} r_{21}$$

$$a^2{}_1 = (r_{12} r_{21}) / a^2{}_2$$

Limitations

Establishing number of factors underlying in a data set has certain limitations, and it is not always possible to ascertain the factors without additional assumption.

(a) When there are more than two common factors the exact configuration of the loadings cannot be ascertained without the additional assumptions.

(b) The rank theorem applies only when the usual operations (the rules for combining factors to create variables) meet a certain set of conditions.

(c) The rank theorem applies only when the data set is not contaminated by the sampling and measurement errors.

(d) Sometimes even without any sampling and measurement errors the data set may not fit to any factor model exactly. Hence the rank theorem even though provides certain indicators of having underlying factors in the data set but this is not an accurate method to find out the factors.

Rotation of Factors

In the factor analysis when the factor loadings have negative signs the correlation matrix and the adjusted correlation matrix pose certain problems. The main objectives of rotation are:

1 Elimination of the negative loadings.

2. Analysis of data with fewer non-zero loadings, as far as possible.

Method of Rotation

First the variables are plotted according to F_1 and F_2 loadings in the x, y plane. The orthogonal axes are taken to represent the factors and the coordinates of the points are the factors loadings.

Figure 6.4 presents a factor loading matrix in which the numbers are test points of the variables. The x axis measures F_1 and the y axis F_2. Of the 6 variables 3 are in the negative field because 3 factor loadings of F_2 are in the negative sign. In order to eliminate the negative signs, the procedure is to select one test point which is further away down in the negative field. We find in the Figure that the test point No. 4 (X_4) is further away in the negative field, hence this point is selected as the basis for rotation. Draw a straight line connecting this point T to the origin of the diagram. Measure the angle F_1 OT; which has been calculated as 22°. Find out Sign theta and cosign theta for 22°, which are respectively: Sign $\theta = .37$ and Cos. $\theta = .93$. Then post multiply the factor loading matrix:

$$\begin{bmatrix} \text{Cos. } \theta & \text{Sin. } \theta \\ -\text{Sin. } \theta & \text{Cos. } \theta \end{bmatrix}$$

The calculation of rotated factor loadings are shown in Table 6.18.

Table 6.18: Rotation of Factor Loadings

Unrotated Factor loadings *Orthogonal Matrix* *Rotated Factor loadings*

$$\begin{array}{cc} F_1 & F_2 \\ \left[\begin{array}{cc} .55 & .15 \\ .43 & .22 \\ .25 & .13 \\ .53 & -.22 \\ .48 & -.13 \\ .25 & -.06 \end{array}\right] \end{array} \text{X} \begin{bmatrix} .93 & .37 \\ -.37 & .93 \end{bmatrix} = \begin{array}{cc} F_1 & F_2 \\ \left[\begin{array}{cc} .45 & .34 \\ .32 & .36 \\ .18 & .21 \\ .57 & .0 \\ .50 & .06 \\ .25 & .03 \end{array}\right] \end{array}$$

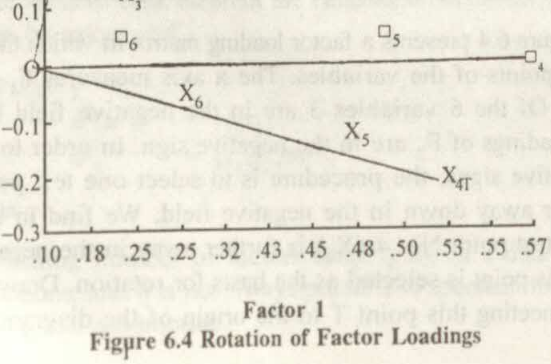


Figure 6.4 Rotation of Factor Loadings

The line in the x-y positive plane OR forms 90° to OT of test 4 or X_4. After the rotation, OT merges with x axis and OR merges with y axis. The new factor loadings after rotation are shown as 'square sign' in figure. In factor analysis one is interested in the influence of factors on the observed variables, hence in the rotation process negative/zero signs are eliminated and the factor loadings are proportionately reduced without effecting the communality. It can be noted from Table 6.19, that even though the factor loadings are changed, and all the negative signs become positive, the communalities of both rotated and unrotated factor loadings remained the same.

Table 6.19: Communalities in a Rotated Factor Loadings

Unrotated		*Communality* h^2		*Rotated*
$(.55)^2 + (.15)^2$	=	.32	=	$(.45)^2 + (.32)^2$
$(.43)^2 + (.22)^2$		.23		$(.32)^2 + (.36)^2$
$(.25)^2 + (.13)^2$		.08		$(.18)^2 + (.21)^2$
$(.53)^2 + (-.22)^2$		.33		$(.57)^2 + (0)^2$
$(.48)^2 + (-.13)^2$		.25		$(.50)^2 + (.06)^2$
$(.25)^2 + (-.06)^2$		.06		$(.25)^2 + (.03)^2$

Factor loadings can be rotated by taking a positive test point to eliminate the smaller loadings and to condense the loadings proportionately.

7

Discriminant Analysis

Discriminant analysis was developed by R.A. Fisher in 1936 who was investigating to solve certain problems of physical anthropology and biology. In social sciences the analysis was first applied by M.M. Tatsuoka and D.V. Tiedeman in 1954 for psychological and educational testing of children. Since then psychologists have been using this technique extensively for personnel and educational testing.

Political scientists can study the voting behaviour of the citizens or legislators with the help of discriminant analysis. By isolating variables that discriminate citizens, one can predict voting pattern of the citizens for ruling party versus opposition party. Anthropologists can group a skull found in an excavation to a particular period through the technique of discriminant analysis. Doctors and physicians use this technique to find out the effects of medical treatment.

Discriminant analysis is an useful tool for the economists too. With the help of this technique one can ascertain the economic differences of two regions or two states, or among states for evolving a suitable strategy for development. Market behaviour among different groups and nature of consumption among different groups of consumers etc., can be predicted with the help of this technique.

What is Discriminant Analysis?

Discriminant analysis is a statistical technique to study the differences between two or more groups of objects with respect to several variables simultaneously. The basic assumption is that there are two or more groups exist and it is presumed that they differ on several characteristic variables. These variables are called discriminant variables, which can be measured at the ratio or interval levels. Discriminant analysis helps a researcher in two ways:

1. It establishes the degree of differences between and among groups.
2. It provides with a means to classify any case into a group with which it most closely resembles.

In other words discriminant analysis is used for the interpretation of group differences and also classification of a particular case into a group. In simple, the discriminant analysis establishes relationships among groups in terms of discriminating variables. Hence assigning an observation X of unknown origin to one of the two or more distinct groups on the basis of the value of the observation is the main technique of the discriminant analysis.

In studying the differences between two or more than two groups of objects, two of the main steps in this analysis are:

1. First to find out the main variables which are responsible to distinguish groups.
2. Second, combination of discriminant variables into an equation to know the magnitude of the differences, which is known as discriminant function. It is a mathematical equation which combines group characteristics to identify a group.

Main Use of the Analysis

The two main uses of discriminant analysis are:

1. ***Interpretation of group differences:*** On the basis of a certain characteristics of a group, the groups are discriminated. The main enquiry relates to find out the magnitude of differences. The characters used to distinguish among the groups are called the "discriminant variables".
2. ***Classification of cases into group's:*** On the basis of discriminating functions of the groups and utilising discriminating variables the cases are classified into groups.

In discriminant analysis it is assumed that the data cases are mutually exclusive. In other words each case must fit into a group. The cases which cannot be included in any group are treated as unclassified cases.

Limitations

The discriminant analysis operates under the following conditions:

1. A group is drawn from a population which has a multivariate normal distribution. This exists when each variable has a normal distribution about fixed values on all others. This enables to compute precisely the tests of significance and probabilities of group members.
2. Population co-variance matrices are assumed as equal for each group. Accordingly the linear discriminant function is a simple linear combination of discriminating variables.
3. When two variables are perfectly correlated both cannot be used at the same time in the discriminant analysis. Hence only such variables are taken as discriminant variables which are not perfectly correlated.
4. When the variables are selected for the discriminant analysis one should see that no variable is a linear combination of the other discriminant variables.

Under discriminant analysis when the variables are analysed to identify their groups, it should be noted that the groups are not dependent on the discriminant variables, if they are, then the analysis becomes a multiple regression.

Mathematical Model

The discriminant analysis can be divided into three components: Groups, Cases and Discriminant Variables.

1. ***Groups:*** In a discriminant analysis the minimum number of groups must be two. That is $g \geq 2$, where g denoted as group.
2. ***Cases:*** A group is composed of number of cases. A case is identified with a group on the basis of its discriminant variables. The cases are denoted as

n_i : Number of cases in group i.

n. : Total number of cases in all the groups.

3. ***Discriminant Variables:*** Each case is composed of certain characteristics, which are known as Discriminant Variables. In discriminant analysis these variables are responsible for identification of a case with a group in the data set. Any number of discriminant variables may be taken for analysis but there is an imposition of limit. The number of discriminant variables in a study must be less than the total number of cases minus two; that is

$$0 < p < (n.-2)$$

where p is the number of discriminant variables. The discriminant variables are indicated as

$$x_1, x_2, \ldots x_p.$$

The relationship between groups, cases and discriminant variables are. In the figure there are three groups and seven cases. The first group has three cases while second and third group has two cases each. The number of discriminant variables are 5 which are shown as dots in each case.

Discriminant Function

A discriminant function is a linear combination of the discriminating variables, which are formed to satisfy certain conditions. The number of functions required for the analysis is determined on the total number of groups. The number of functions are one less than the total number of groups. For instance, if there are four groups, then the number of functions are three. Generally more variables are taken than the number of groups. In the event of less number of variables than the number of groups, number of unique functions are determined as number of variables minus one. In other words the number of unique factors are determined on the formula of number of groups minus one, or number of variables minus one, whichever is the lowest. When there are 5 groups and 3 variables the number of functions taken for analysis are 2, i.e. (3–1) which is the lowest.

Derivation of Co-efficients: The nature of the function is that if there are three functions, the first function indicates the maximum differences among the groups in the analysis, and the second function lesser magnitude of differences among the groups, and the third still lesser magnitude than the second; and more the functions less and less are the magnitude of the function in indicating the group differences.

After the first function, for the subsequent functions there is another condition to be remembered. The values of the second function are assumed to be uncorrelated with the values of the first function. Accordingly the coefficients for the subsequent functions are defined which maximises group differences but remain uncorrelated with the previous functions.

The discriminant Function is stated as follows:

f_{km} $= U_0 + U_1X_{1km} + U_2X_{2km} + U_3X_{3km} \ldots\ldots + U_pX_{pkm}.$

f_{km} : Value of the discriminant function for case m in group k. This is also called discriminant score.

U_0 : $\overset{p}{\underset{i=1}{\varepsilon}} U_iX_i\ldots$ ($X_i\ldots$ is the mean value of variable i for all cases).

U_i $= U_i\sqrt{n.-g}$

U_i : It is the coefficient which produces the desired characteristics in the function.

X_{ikm} : Value on discriminating variable X_i for case m in group k.

When Us are employed the value of discriminating scores for the data set are in standard form. This makes the discriminating scores over all the cases as mean zero and within groups standard deviation as 1. The discriminating score for a given case represents the position of that case along the same axis, defined by the function. If there are two functions, and as they are orthogonal, they can be plotted in a x, y plane to locate the position of the case.

Total Matrix

Total or T matrix is a total sum of squares and cross products matrix. This is a square symmetric matrix. In other words it is a total dispersion matrix. It summarises how much the points are spread out around the total space defined by all the variables dispersion. Each element in the T matrix is computed as:

$$t_{ij} = \sum_{k=1}^{g} \sum_{n=1}^{n_k} (x_{ikm} - x_i \ldots)(x_{jkm} - x_j \ldots)$$

The terms in the parentheses are the amount by which the value of a particular case deviates from the grand mean.

$x_i..$: This is the grand mean, that is, the mean value of variable i for all cases. In a symmetrical matrix where i = j, the two terms in the parenthesis are the same; hence it is only squaring the deviations. Each diagonal element is the sum of squared deviations from the grand mean.

The T matrix can be converted into a co-variance matrix or into a matrix of coefficient of correlations on the basis of the following formulae:

1. Co-variance: By dividing each element of T matrix by (n–1) a T matrix can be converted into a co-variance matrix. [t/(n–1)]
2. Correlation: By dividing each element of a T matrix by the square root of the product of the two diagonal elements falling in the same row and column, a T matrix can be converted into a correlation matrix. For example, to find out r_{12} the formula accordingly is

$$r_{12} = \frac{t_{32}}{\sqrt{(t_{11})(t_{22})}}$$

Within Group Matrix

The within groups sum of squares and cross products matrix is denoted as W matrix. In this matrix deviations are measured from the mean of the group to which the case belongs instead of from the grand mean as in case of T matrix. Each element of W matrix is then:

$$w_{ij} = \sum_{k=1}^{g} \sum_{m=1}^{n_k} (X_{ikm} - \overline{X}_{ik})(X_{jkm} - \overline{X}_{jh})$$

Where $\overline{X}_{ik}$ and $\overline{X}_{jk}$ are mean from respective groups.

A within group matrix can also be converted into a co-variance or correlation matrix. In case of correlation the formula adopted for T matrix can be used for the W matrix for conversion. That is, by dividing each W_{ij} by square root of the product of two diagonal elements falling in the same row and column. For conversion into a co-variance matrix the formula is:

$$\text{Covar. } C_{ij} = W_{ij} / (n-g)$$

Centroids

In a discriminating space two types of centroids can be located; grand centroid and group centroids. Grand centroid is that point in space where all of the discriminating variables have their average values over all cases. It is the centre position for all points representing the data cases.

A group centroid is an imaginary point which has the co-odrinates that are the group's mean on each of the variables. Each centroid represents the typical position of its group. Hence by looking the group centroids one can know how the groups differ. The value of discriminant function represents the co-ordinates in the space of discriminant function. By looking into a groups centroids one can tell where it is located relative to the grand centroid or the centre of the system. The distance between the grand centroid and the group centroid is measured in discriminant scores in terms of standard deviation units. A group represents the number of standard deviations it is from the grand centroid.

Between Groups Matrix

When there are no differences among the group centroids, the T matrix is just equal to W matrix element by element. That is T = W. If the groups centroids are different, the elements in the W matrix will be smaller than the corresponding elements of T matrix. Hence:

$$T - W = B$$

This difference matrix B is called between groups sums of squares and cross product matrix. The size of the elements of B in relation to T gives a measure that distinguishes the groups.

Calculation of v: For the linear discriminant functions the calculation of U's are important. The calculation of U is based on v; hence the calculation of v's are the main concern of discriminant analysts. The elements of W and B can be utilised for the calculation of v's with one condition that the sum of the squared values of v's must be equal to 1. In the following equations b and w are the known quantities calculated from the sample data. So, the purpose is to solve these simultaneous equations for λ (lamda) and v's.

$$b_{1i}v_i = \lambda w_{1i}v_i$$
$$b_{2i}v_i = \lambda w_{2i}v_i$$
$$b_{3i}v_i = \lambda w_{3i}v_i$$
$$b_{pi}v_i = \lambda w_{pi}v_i$$

where λ is a constant called eigen value, and vi a set of p coefficients. Each solution yields its own lamda and set of v's corresponding to one discriminant function. On this basis for each case according to number of functions the discriminant scores can be computed.

Discriminant Scores

Table 7.1 gives B.A. Bardes's data with regard to senatorial voting pattern which he analysed through discriminant analysis in 1973. The three functions define a three dimensional space in which the case can be located. The first function can be the horizontal axis, the second function which is uncorrelated to the first, hence orthogonal and can be the vertical axis; and the third function perpendicular to both, as it is uncorrelated to both the functions.

According to the discriminant function formula the score is computed by taking the original values for a case on each variable and multiplying it by the coefficient for that variable and addition of these products along with the constant term. The constant term is the adjustment for the mean discriminant score that will be zero over all the cases.

If there are small number of cases, then case by case inspection is possible, but if there are large number of cases the group centroids have to be located for comparison, because the group centroid is a most typical position for the group. A case's position nearest to a

Table 7.1: Computation of Discriminant Scores for a Case

Variable	*Function 1*			*Function 2*			*Function 3*		
	Coef.	*con.*	*Coef x Con.*	*Coef.*	*con.*	*Coef. x Con.*	*Coef.*	*con.*	*Coef x Con.*
Constant			5.4243			3.5685			–4.3773
X_1	.8078	1	.8078	–.5225	1	– 5225	1.6209	1	1.6209
X_2	.7940	3	2.3820	–1.1177	3	–3.3531	–.3339	3	–1.0017
X_3	–4.6004	1	–4.6004	–1.1228	1	–1.1228	–1.1431	1	–1.1431
X_4	–.6957	3	–2.0871	–1.3160	3	–3.9480	1.1418	3	3.4254
X_5	–1.1114	1	–1.1114	1.1132	1	1.1132	.3871	1	.3781
X_6	1.4387	1	1.4387	1.0422	1	1.0422	.2000	1	.2000
Discriminant Score			2.2539			–3.2225			–0.8977
or			2.25			–3.22			–0.90

Source: Klecka, p. 24 ff.

group's centroid can easily be known and the case can be included in the group.

If there are two functions the position of the cases can be shown in a graph, by measuring the first function in the x axis and the second discriminant function in the y axis. A two function, three group and 20 case hypothetical data is plotted in Figure 7.1 for visual inspection. The small zeros are the cases and the number indicate the group to which it belongs. Here there are three distinct groups but two cases m_1 and m_2 are far away from these groups and cannot be identified with any of the groups. The asterisks inside the groups indicate the group centroids.

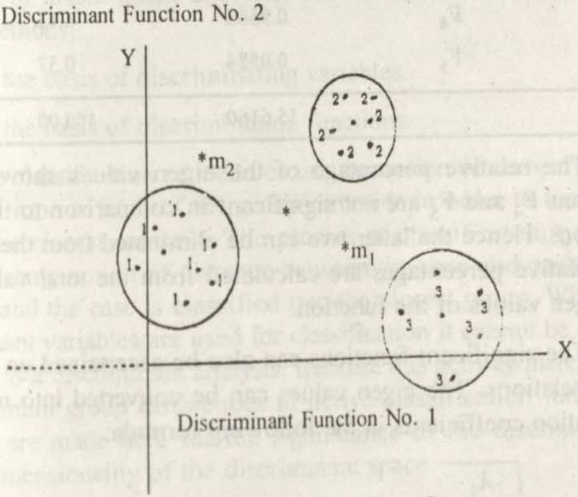


Figure 7.1 Plotting of Cases to form Groups

Significant Functions

According to the number of groups and the variables the number of discriminant functions are determined. But all the functions may not be taken for analysis, only significant functions are retained for the groupings. The significant functions are selected on the basis of eigen values. As explained earlier, a set of v's and lamda or the constant called eigen value can be derived for the each discriminant function. These eigen values are the indicators for the selection of functions. The lamdas are either positive or zeros. Higher the eigen value more powerful is the function as a discriminator. Accordingly the functions with smallest eigen values are eliminated from the study.

Eigen values for four functions from a hypothetical data are presented in Table 7.2 to evaluate their importance.

Table 7.2: Selection of Significant Functions

Function	*Eigen value*	*Relative Percentage*
F_1	8.5328	54.64
F_2	3.6821	23.58
F_3	2.3864	15.28
F_4	0.9563	6.13
F_5	0.0584	0.37
	15.6160	100.00

The relative percentage of the eigen values show that the functions F_4 and F_5 are not significant in comparison to the earlier functions. Hence the latter two can be eliminated from the analysis. The relative percentages are calculated from the total value of all the eigen values of the function.

The significant functions can also be ascertained on the basis of correlations. The eigen values can be converted into respective correlation coefficients in the following formula:

$$r_1 = \sqrt{\frac{\lambda_i}{1+\lambda_i}}$$

where λ indicates the eigen values of respective discriminant function. As per the formula the correlation coefficient for the first function is

$$r_1 = (8.5328)/(1+8.5328) = 0.90$$

$$\sqrt{0.90} \quad = 0.95$$

Accordingly correlation coefficients for other functions are $r_2 = 0.89$, $r_3 = .84$, $r_4 = .75$ and $r_5 = .24$. A high coefficient indicates that a strong relationship exists between the groups as shown by the first discriminant function. But the first function may not have the largest cantonical correlation always. The first function may be a most powerful one, but it may so happen that it may be weakly related. If the groups are not very different the relationship will be weak. Hence one has to examine both relative percentage and the correlations in order to determine the number of significant discriminant functions for the study.

Classification of Cases

Classification of cases into different groups is another important objective of discriminant analysis. The classification can be done in two methods:

1. On the basis of discriminating variables.
2. On the basis of discrimination functions.

The classification procedures use the notion of comparing the cases' position to each of the group centroids in order to locate the case to the closest centroid. The process of classification typically involves some notion of distance between the case and each group centroid and the case is classified into the closest group. When the discriminant variables are used for classification it cannot be strictly regarded as a discriminant analysis. Because this activity merely uses the maximum group differences to derive classification functions. No tests are made here for the significance of the discrimination on the dimensionality of the discriminant space.

Simple Classification Function

In 1936, Fisher suggested a classification function based on linear combination of the discriminant variables. This linear combination maximises group differences while minimising variations within the groups. A classification function for each group can be computed on the following formula:

$$h_k = b_{k0}+b_{k1}x_1+b_{k2}x_2+b_{k3}x_3.......+b_{kp}x_p.$$

Where h_k is the score for the group k, and the b's are the coefficients which have to be derived. A case is classified into a group with highest score of h. The coefficients are calculated as under:

$b_{ki} = (n - g) \sum_{j=1}^{p} a_{ij}x_{jk}.$

b_{ki} : It is the coefficient for variable i in the equation corresponding to group k.

aij : It is an element from the inverse of the within groups sum of squares and cross product matrix, or w^{-1}.

A constant term is required for equation and it is defined as

$$b_{k0} = \sum_{j=1}^{p} b_{kj}x_{jk}.$$

where k is the mean value of variable j for those cases in group k. After getting the coefficients the actual data of the variables x_1, x_2, etc., are used to compute h_k.

Generalised Distance Functions

In this method the classification is done on the basis of the measurement of distances from the individual cases to each of the group centroids and a case is classified into the closest group. Mahalnobis in 1963 defined a generalised distance measure which is suitable to solve the problem.

The formula is:

$$D^2 (X/G_k) = (n. - g) \sum_{j=1}^{p} \sum_{j=1}^{p} a_{ij} (X_{jk} - X_{ik}) (X_j - X_{jk})$$

D^2 (X/G_k) is the squared distances from the point X of the location of the case to the centroid of group k. After the calculation of D^2 the case is classified into the group with smallest D^2. The D^2 has same properties as Chi square statistic with p degrees of freedom. The distance between a case and the group's centroid is measured in Chi square units. This is because it is assumed in the analysis that each group comes from a population with a multivariate normal distribution. It is found that, when plotted in a graph, most of the cases are clustered near the centroid, and the density of the cases diminish in a precise fashion as they get further away from the centroid.

Classification with Discriminant Functions

Instead of discriminant variables, the discriminant functions can also be used for the purpose of classification of a case. The simple classification formula of the variables are used, but instead of the variables, f's are used. Hence the formula is:

$$h_k = b_{k0} + b_{k1}f_1 + b_{k2}f_2 + b_{k3}f_3$$

Since the number of discriminant functions are smaller in number than the discriminant variables, this method is a simpler one. But under this classification the groupings may be different because of two factors:

1. When the group covariance matrices are not equal the classification may be different. This is because the procedure deriving the discriminant function, within group cross product matrix is used which is a weighted average of the individual group cross product matrix.
2. The classification may be different if one or more of the discriminant functions are ignored in the analysis because they are not statistically significant. Even though the results may be different in this procedure, but the classification on the basis of discriminant functions is more accurate since in this method the effects of idiosyncratic sample variations are reduced.

8

Cluster Analysis

Classification of data is an important aspect in research. Data have to be classified according to the need of the research design for analysis. The study of the science of classification is of recent origin and known in various names such as typology and taxonomy. Now the science of classification is termed as cluster analysis. Indians are the first analysts of classitication. Muni Bachhayana, centuries earlier to Christ, was the earliest user of cluster analysis in his book *"Kama Sutra"* or Principles of Love. He classified men and women into four clusters on the basis of their physical structure, mental horizon and social behaviour in order to analyse the love-life of human beings. He classified men into four classes; *Aswa* (horse), *Brisha* (bull), *Mriga* (deer) and *Sasa* (hare) in descending order of strength and beauty. In classifying females Bachhayana has not compared them with the animals as in case of males except the last class. The four clusters of females are: *Padmini* (lotus, the symbol of tenderness, beauty and fragrance) *Chitrini* (beauty as if a painting) *Sankhini* (smooth and white as conch shall as well as sweet voice) and *Hastini* (female elephant).

In modern times J. Czekanowski, a German anthropologist, formalised the principles of classification for cluster analysis. In 1930s psychologists, specially J. Zubin and R. Tryon further developed the procedures. The latter's publication of *"Cluster Analysis"* in 1939 firmly established the cluster analysis as an important tool for classification of entities. The cluster analysis has been widely utilised in biological and social sciences from 1960 onwards especially after the publication of two important works: *"Principles of Numerical*

Taxonomy" by R. Sokal and F. Rohlf in the year 1963 and *"Pattern of Clustering by Multivariate Mixture Analysis* "by J.M. Wolfe in 1970.

Cluster analysis is a technique to group variables, individuals and entities. Once the variables are classified on certain characteristics it makes the work of researcher easier for further analysis by taking only a smaller sample space. Otherwise once the species of an entity is classified, it is easier to analyse it, since the characteristics of the class to which it belongs is already known. Cluster analysis is regarded as an alternative for factor and principal component analysis for the reduction of data in a research design.

A cluster, according to B. Eviritt is a "continuous regions of space containing a relatively high density of points separated by such other regions by containing a relatively low density points." In other words cluster is a homogenous group of cases, variables or entities. Identification of clusters inherent in a data set is called cluster analysis.

The most important use of cluster analysis is development of a typology or classification. In economics, especially in applied economics, where an analysis is made on the data collected from the primary sources, classification is a first and the most important aspects of research. Classification of households on the basis of socio-economic-cultural parameters is a good basis for analysing consumer behaviour. Hence, the cluster analysis is regarded as the first step for any empirical study in economics.

Investigation of useful conceptual schemes for grouping variables or entities is another use of cluster analysis. In different disciplines different data sets require special procedures for the classification of the variables according to the need of the researchers. Accordingly a researcher has to explore different methods of classification or clustering that can be applied in other cases of research in similar situations.

Thirdly, hypothesis generation and hypothesis testing is another field of cluster analysis. Analysis of data to generate hypothesis with regards to the nature of classification is a research in specific field. Testing can be taken also for certain hypotheses in order to determine whether the hypothesis is present or not in the data set as already defined through other procedures.

Importance of cluster analysis has increased in recent years in inorganic chemistry, atomic theory of matter and in science of medicine because in these fields classification contains major field of research. Now, in social sciences the cluster analysis is widely used because of the availability of high speed computers which can handle large matrices to analyse field data.

Steps in Cluster Analysis

There are four basic steps through which a cluster analyst can work to classify entities.

1. ***Selection:*** The first stage of the analysis is to select a suitable sample size from the problem population on which the cluster analysis will be made.
2. ***Definition of variables:*** The variables in the data set has to be defined clearly so that the entities in the sample can be measured.
3. ***Computation of similarities:*** In the third stage the entities have to be measured for their closeness in terms of similarity or dissimilarity measures. In cluster analysis separate formulas have been devised to find out the closeness of the entities on binary and other procedures. A researcher has to select one of the methods for cluster analysis.
4. ***Groupings:*** Grouping is the final stage of the cluster analysis. There are several methods of grouping in cluster analysis. One popular method adopted by the researchers is to construct dendrograms to identify the number of cluster in the entities.

Similarity Measures

Classification of entities require investigations to search for identical characteristics or similarities among the entities and that should be measured quantitatively. In general analysis the correlation coefficient is regarded as the most important method of similarity measure to find out the degree of closeness. Karl Pearson's product moment correlation coefficient is widely used by the researchers to find out the correlation. But the correlation has not been accepted as a good measurement of similarity due to its certain drawbacks:

1. One important drawback of correlation is that it is sensible to the shape. Two profiles can have a perfect correlation of +1, yet they may not be in identical ones.
2. A correlation is insensitive to differences in the magnitude of variables. Two sets of data may have same correlation but in one set the quantities may be small and in the other the quantities are at higher level.
3. The two data sets may scattered differently but the correlation may be the same. Here one has to workout their respective dispersion in order to compare the data sets.
4. When the correlation coefficient 'r' is calculated on the basis of linear relationship some information is lost and a portion of entities remain unexplained.
5. The correlation fails to satisfy the triangular inequality, correlation of cases does not make statistical sense because one must obtain the mean value across different variables, rather than across cases.

The Concept of Similarity

The quantitative estimation of similarity is dominated by the concept of matrices. It represents cases as points in a coordinate space such that the observed similarity or dissimilarity of the points corresponding to the matric distances between them.

There are four standard criteria to judge a similarity measure. The are:

1. ***Symmetry:*** If there are two entities x and y; and d denotes as distance; distance between them is expressed as:

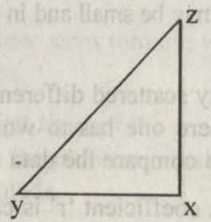
   $d(x, y) = d(y, x) \geq 0$

A numerical function d (x, y) of pairs of points of a set E is said to be matric for E, when it satisfies the above condition or criterion.

2. ***Triangular Inequality:*** If there are three entities x, y, and z and distance between any two entities is less than or equal to the sum total of the distance between the two other pairs of entities it is called triangular inequality.

   $d(x, y) \leq d(x, z) + d(y, z)$

   Since the three entities are in the three corners of a triangle and their distances are measured and compared this criterion is known as triangular inequality.



3. ***Distinguishability of non-identicals:*** If distance between x and y is not zero, then x is not equal to y.

   If $d(x, y) \neq 0$ then $x \neq y$

4. ***Indistinguishability of Identicals:*** If two entities are identical say x and x′ ; then their distance is zero.

   $d(x, x') = 0$

Similarity Coefficients

Computation of similarity coefficients is the first step in cluster analysis. When the data are in nominal level the similarity coefficients can be calculated on a binary scale and can be presented in a matrix form. Computation of similarity coefficients through binary character is easier since in a binary scale the enquiry is confined to whether an entity has certain character or not. If a character is present (or answer is yes) it is stated as '1' or otherwise (if answer is no) as '0'. These 1s and 0s are used, that is only two figures (binary), to compute similarity coefficients between the two entities.

Example: Say we have to make a classification of households on the basis of their consumption behaviour in a locality for which the families are divided into two categories of middle-income group (M) and lower-income group (L). It is further assumed that there are three types of families reside in the locality viz., schedule castes (S), schedule tribes (T) and the general caste (G). Information obtained from 20 sample households on the consumption behaviour of 8 items of consumer goods and durables are presented in Table 8.1.

Three formulas to calculate similarity coefficients based on binary data generally used in cluster analysis are described below with the help of sample binary data from Table 8.1.

1. ***Simple Matching Coefficients:*** For qualitative data based on binary system are calculated in terms of 2 × 2 association tables as explained below:

		Individual i	
		1	0
Individual j	1	a	b
	0	c	d

Table 8.1: Sample Data for the Household Consumer Behaviour

	Case	Consumer items							
		Pan	*Cigarette*	*Non-veg.*	*Wheat*	*Gas*	*Scooter*	*TV*	*Own House*
		(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*	*(8)*
1	MS	1	1	1	0	0	0	1	0
2	MG	0	1	1	1	1	0	1	0
3	MS	0	1	1	0	0	0	0	0
4	MS	1	1	1	0	0	0	1	0
5	MS	0	1	1	1	1	0	0	1
6	MG	0	1	0	1	1	1	1	0
7	MG	1	0	1	1	1	0	1	0
8	MT	0	1	1	0	0	0	0	0

(Contd.)

	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*	*(8)*
9 MT	0	1	1	1	0	0	1	0
10 MT	1	0	1	0	0	0	0	0
11 LS	0	0	1	0	1	0	1	0
12 LS	1	1	1	1	0	0	0	0
13 LS	0	1	1	0	0	0	1	0
14 LT	0	1	1	1	0	0	1	1
15 LT	0	0	1	0	0	0	1	1
16 LG	1	0	0	1	1	1	1	1
17 LG	0	1	0	1	1	0	1	0
18 LG	0	0	1	0	0	0	1	1
19 LG	0	1	1	1	1	1	1	0
20 LG	0	0	0	1	1	0	1	1

The Simple Matching Coefficient Formula is

$$S = \frac{(a+d)}{(a+b+c+d)}$$

S is the similarity measure between the two cases which ranges from 0 to 1. This can be easily transformed into a matrix. Let us take two cases from Table 8.1 to calculate the coefficient as under. The two cases are:

MS : 1 1 1 0 0 0 1 0 (1)
MG : 0 1 1 1 1 0 1 0 (2)

From the above data the a, b, c, d, data calculated are: a = 3, b = 2, c = 1, and d = 2. Accordingly the coefficient is:

$$S = \frac{(3+2)}{(3+2+1+2)} = 0.625$$

2. ***Jaccard's Coefficient:*** Even though based on binary data, Jaccard's Coefficient is little different from the simple matching coefficients. His formula denoted as J is

$$J = \frac{a}{(a+b+c)} = 3/6 = 0.5$$

In Jaccard's formula the d cell which is 00 has been eliminated, hence the Jaccard's coefficients are lower to the simple matching coefficients.

In Table 8.2 similarity coefficients of 5 cases in matrix form are shown, in which the upper triangle shows the simple matching coefficients and Jaccard's coefficients in the lower triangle. It can be seen that under simple matching there are more ties among the coefficients while in the Jaccard's coefficients the number of ties are very much less.

Table 8.2: Simple Matching and Jaccard's Coefficient

Entity	MS	MG	MT	LS	LT
MS	—	.50	.55	.50	.63
MG	.43	—	.63	.75	.75
MT	.50	.40	—	.63	.63
LS	.33	.60	.25	—	.50
LT	.50	.67	.40	.33	—

Apart from the two above formulas some other formulas are also used by the cluster analysts in which the d cell is avoided. Three such formulas are given below:

$$(1)\quad J_1 = \frac{2a}{a+b+c} = \frac{6}{6+2+1} = 0.666$$

$$(2)\quad J_2 = \frac{a}{a+2(b+c)} = \frac{3}{3+4+2} = 0.333$$

Two of the other formulas in which d cell is associated are

$$S_1 = \frac{2(a+d)}{2(a+d)+b+c} = \frac{6+4}{6+4+2+1} = 0.769$$

$$S_2 = \frac{a}{a+b+c+d} = 3/8 = 0.378$$

3. ***Grower's General Similarity Coefficient:*** Prof. J. Grower used binary data for both qualitative and quantitative variables for cluster analysis in biometrics. His formula is:

$$S_{ij} = \sum_{k=1}^{p} S_{ijk} / \sum_{k=1}^{p} W_{ijk}$$

S_{ijk} : Similarity score based upon the outcome of the comparison of variable k across i and j

W_{ijk} : w is the weighting variable.

Scoring System

Rules for S_{ij} : 1, 1 = 1; 0, 1 = 0, 1, 0 = 0 and 0, 0 = 0

Rules for W_{ij} : 1, 1 = 1; 0, 1 = 1; 1, 0 = 1 and 0, 0 = 0

Let us work out an example from Table 8.1

Case i (5) MS:	0	1	1	1	1	0	0	1	
Case j (10) MT:	1	0	1	0	0	0	0	0	
$S_{ijk} =$	0	0	1	0	0	0	0	0	= 1
$W_{ijk} =$	1	1	1	1	1	0	0	1	= 6

S_{ijk}/W_{ijk} = 1/6 = 0.17

This formula's result is almost similar to Jaccard's coefficients. In case of qualitative data Grower preferred simple matching coefficients with a slightly different from the simple matching formula explained earlier. Grower's scoring for simple matching coefficients is S_{ijk} computed as 1, 1 = 1; 0, 0 = 1, and rest of the associations i.e. 1, 0 and 0, 1 as equal to 0. The formula is,

$$S_{ijk} = \frac{(a+d)}{(a+b+c)}$$

For the same data of cases 5 and 10, of Table 8.1, the calculations can be made afresh.

S_{ijk} :	0	0	1	0	0	1	1	0	= 3
W_{ijk} :	1	1	1	1	1	0	0	1	= 6

Hence $\sum S_{ijk} / \sum W_{ijk} = 3/6 = 0.5$. This value is a higher one compared to the earlier coefficient computed as per the previous scoring.

Quantitative Data

Grower provided a separate formula for the computation of similarity coefficients for the quantitative data. The formula is,

$S_{ijk} = 1 - x_{ik} - x_{jk} \; 1/R_k$

X_{ik} : score of individual i on variable k.

X_{jk} : score of individual j on variable k.

R_k : Range of variable k.

In Table 8.3 height and weights of 5 individuals are given. We can find out similarity coefficients according to Grower's formula for quantitative data.

Let us calculate the similarity coefficient between the serial No.1 and 2. The range of height is 10 (70–60) while the range of weight is 9 (51–42)

$$
\begin{aligned}
\text{Height : } S_{ijk} &= 1 - |\,(60 - 62\,|\,/10 \\
&= 1 - (2/10) \\
&= 1 - 0.2 \\
&= 0.80
\end{aligned}
$$

Table 8.3: Heights and Weights of Students

Sl. No.	*Heights (Inches)*	*Weight (Kgs.)*
1	60	45
2	62	50
3	65	45
4	67	24
5	27	51

$$
\begin{aligned}
\text{Weight: } S_{ijk} &= 1 - 45 - 50/9 \\
&= 1 - (5/9) \\
&= 1 - 0.56 \\
&= 0.44
\end{aligned}
$$

Hence the similarity coefficient between the two in terms of heights and weights is

$$S_{ij} = \frac{0.80 + 0.44}{2} = 0.62$$

On the basis of the formula, similarity coefficients can be found between any two individuals taking both height and weights into account and shown in a matrix form. If the variables have more than two characteristics accordingly their mean score can be computed as similarity coefficient.

Distance Measures

Distance measure is another method through which a cluster analysis can be developed for a set of data. The main characteristics of distance measure are as follows:

1. In a similarity measure clusters are formed in terms of their similarity but in distance measure the clusters are separated on the basis of their distances or dissimilarities.
2. Distance measure is just the reverse of similarity measure. But in a similarity measure, as in case of a correlation coefficient the values remain within 0 and 1 while a distance measure can take any positive value. Here there is no upper bound of the value.
3. One difficulty with the distance measure is that it is scale dependent. The variables may have in different values and scales. If the raw data is converted into a standardised form the distance measure is more suitable for clustering in the entities. It can be standardised with the help of the following formula:

$$Z_{ik} = \frac{(X_{ik})}{k}$$

k is the standard deviation of *k*th variable.

Euclidean Matric Measure

Euclidean formula for the distance measures most popular in forming clusters by the researchers. The formula is:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (X_{ik} - X_{jk})^2}$$

or simply the squared Euclidean distance as:

$$d^2{}_{ij} = \sum_{k \neq 1}^{p} (X_{ik} - X_{jk})^2$$

Where X_{ik} and X_{jk} are the value of *k*th variable in the *i*th & *j*th case respectively.

A distance function can be transformed into similarity function in the following formula but the reverse is regarded as the most difficult one.

$S_{ij} = 1/(1 + d_{ij})$

Grower explains that a similarity function can be transformed into a distance measure provided that the matrix of similarity coefficients are all positive and semi definite. His formula is:

$d_{ij} = 2\,(1 - S_{ij})$

Separation of entities on the basis of Euclidean distance for the formation of clusters provides completely a different information in comparison to clusters based on correlation coefficients which is an analysis of association. In order to show their differences the product moment correlation coefficients and the Euclidean coefficients calculated for the same data with ranks are shown in Table 8.4 (a) and (b) respectively. The upper triangles show the coefficients while the lower triangle indicate the ranks.

Table 8.4(a): Product Moment Correlation Coefficients

	A	B	C	D
A	—	.776	.702	.742
B	(3)	—	.729	.779
C	(6)	(5)	—	.936
D	(4)	(2)	(1)	—

Table 8.4(b): Euclidean Coefficients

	A	B	C	D
A	—	.266	.732	.736
B	(2)	—	.523	.736
C	(5)	(4)	—	.144
D	(6)	(3)	(1)	—

Table 8.5: A Comparison of Ranks

Entities	*Correlation Coefficients*	*Euclidean Coefficients*
C D	1	1
B D	2	3
B A	3	2
A D	4	6
B C	5	4
C A	6	5

In case of the correlation coefficients the ranking is done from the highest to the lowest value in showing similarities while in the distance measures the order is from the lowest to the highest, the opposite of the former. The comparison of their ranks are presented in Table 8.5. It can be seen that except in the case of first one, all the ranks are different in the Euclidean distance measure in comparison to correlation coefficients.

Techniques of Clustering

There are four important methods of clustering in which correlation coefficients, similarity coefficients or dissimilarity coefficients can be taken for forming of clusters.

1. ***Hierarchical Technique:*** The entities are classified at different levels to form a tree or dendrogrum, Greek 'dendro' for tree. It is a visual presentation of clusters. After preparing the dendrogram the acceptable number of classes are determined

for further analysis. This technique is very popular among the cluster analysts.

2. *Optimisation Partitioning Technique:* In this technique clusters are formed by the optimisation of the clustering criteria. Partition is done on the basis of mutually exclusive entities.

3. *Density or Mode Seeking Technique:* In this method clusters are formed on the basis of how densly the entities are spread in the matric space. Dense concentration of the entities are the main basis for the clustering.

4. *Clumping Technique:* Clumping is another method in which clustering is specified. In this method it is assumed that the clusters overlap each other; not exclusive to each other as assumed in other methods of clustering. Specific clusters are extracted from the overlapping of classes.

In social sciences the Hierarchical Agglomerative, Iterative partitioning and factor analytic techniques are widely in use.

Hierarchical Technique

There are two methods in the hierarchical technique: agglomerative method and divisive method. In both the methods clustering can be done on the basis of dissimilarity coefficients arranged in the form of matrices.

Agglomerative Method: The main objective of this method is that it tries to "fuse" the two variables in terms of their dissimilarity coefficients. There are two procedures of fusion:

(1) Single linkage method and (2) Complete linkage method.

Single Linkage Method: Successive fusion of the nearest neighbour is the main function of the single linkage method. The matrix D_1 is a matrix of distances. Each element in the matrix gives the distance between two entities.

Process of this method starts with the selection of the lowest distance between the two entities to fuse them together. In this matrix D_1, 2 is the lowest distance between the entities 1 and 2, hence a the beginning these two variables are fused first.

$$D_1 = \begin{array}{cccccc} & 1 & 2 & 3 & 4 & 5 \\ 1 & 0 & 2 & 6 & 10 & 9 \\ 2 & — & 0 & 5 & 9 & 8 \\ 3 & — & — & 0 & 4 & 5 \\ 4 & — & — & — & 0 & 3 \\ 5 & — & — & — & — & 0 \end{array}$$

On the basis of nearest neighbour principle, the entities (12) are fused with the rest of the entities and D_1 is reduced to matrix D_2 with calculating the cell coefficients afresh.

$$d_{(12)3} = \min (d_{13}, d_{23}) = d_{23} = 5$$

$$d_{(12)4} = \min (d_{14}, d_{24}) = d_{24} = 9$$

$$d_{(12)5} = \min (d_{15}, d_{25}) = d_{25} = 8$$

Accordingly matrix D_2 is

$$D_2 = \begin{array}{ccccc} & (12) & 3 & 4 & 5 \\ (12) & 0 & 5 & 9 & 8 \\ 3 & — & 0 & 4 & 5 \\ 4 & — & — & 0 & 3 \\ 5 & — & — & — & 0 \end{array}$$

The smallest entry in D_2 is element 3 between variables 4 and 5. Hence these two variables have to be fused together. Then the next matrix D_3 is prepared taking the fusion of (45) variables and the linkages are calculated afresh as shown below:

$$d_{(12)3} = \min (D_{13}, D_{23}) = d_{23} = 5$$

$$d_{(12)(45)} = \min (d_{14}, d_{15}, d_{24}, d_{25}) = d_{25} = 8$$

$$D_3 = \begin{array}{cccc} & (12) & 3 & (45) \\ (12) & 0 & 5 & 8 \\ 3 & — & 0 & 4 \\ (45) & — & — & 0 \end{array}$$

In the D_3 matrix the lowest element is 4 between the individuals 3 and (45) and this has to be fused, as per calculation below: The final linkage is shown in D_4 matrix.

$$D_4 = \begin{array}{c|cc} & (12) & (345) \\ \hline (12) & 0 & 5 \\ (345) & — & 0 \end{array}$$

All the fusion values calculated from D_1 to D_4 are tabulated below in Table 8.6. The dendrogram prepared on the basis of this table is shown in Figure 8.1

Table 8.6: Fusion of Dissimilarity Coefficients

Variables Fused	*Coefficient*
1 – 2	2
4 – 5	3
3 – (45)	4
(12) – (345)	5

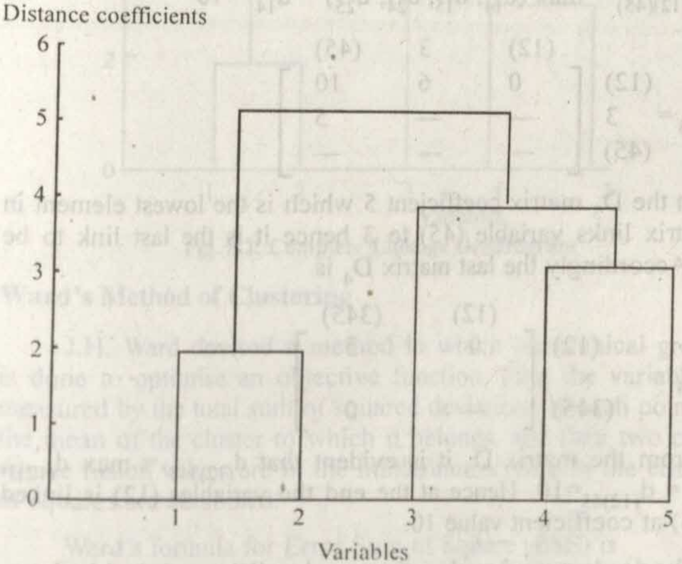


Fig. 8.1 Single Linkage Dendrogram

Complete Linkage Method: Fusion of the variables in this method is done on the basis of "furtherest neighbour" or maximum of the distances but the process is started from the smallest distance first. Accordingly the D_2 matrix is calculated in the following manner.

$$d_{(12)3} = \max(d_{13}, d_{23}) = d_{13} = 6$$

$$d_{(12)4} = \max(d_{14}, d_{24}) = d_{14} = 10$$

$$d_{(12)5} = \max(d_{15}, d_{25}) = d_{15} = 9$$

$$D_2 = \begin{array}{c|cccc} & (12) & 3 & 4 & 5 \\ \hline (12) & 0 & 6 & 10 & 9 \\ 3 & — & 0 & 4 & 5 \\ 4 & — & — & 0 & 3 \\ 5 & — & — & — & 0 \end{array}$$

The lowest element in D_2 is 3 hence variables 4 and 5 are fused. The D_3 matrix is calculated accordingly:

$$d_{(12)3} = \max(d_{13}, d_{23}) = d_{13} = 6$$

$$d_{(12)(45)} = \max(d_{14}, d_{15}, d_{24}, d_{25}) = d_{14} = 10$$

$$D_3 = \begin{array}{c|ccc} & (12) & 3 & (45) \\ \hline (12) & 0 & 6 & 10 \\ 3 & — & — & 5 \\ (45) & — & — & — \end{array}$$

In the D_3 matrix coefficient 5 which is the lowest element in the matrix links variable (45) to 3 hence it is the last link to be fused. Accordingly the last matrix D_4 is

$$D_4 = \begin{array}{c|cc} & (12) & (345) \\ \hline (12) & 0 & 5 \\ (345) & — & 0 \end{array}$$

From the matrix D_3 it is evident that $d_{(12)345} = \max d_{(12)3}$, $d_{(12)45} = d_{(12)45} = 10$. Hence at the end the variables (12) is linked to (345) at coefficient value 10.

The dendrogram based on the complete linkage method is shown in Figure 8.2. It can be verified from the two figures that the lowest fusion value is same for both the simple linkage and the complete linkage dendrograms. But the upper limits are completely different;

the upper limit of complete linkage dendrogram is double of the single linkage dendrogram. In both the cases we find that there are two clusters: 1, 2 in one cluster and 3, 4, 5 in the other, but in case of complete linkage dendrogram the linkages are very clear.

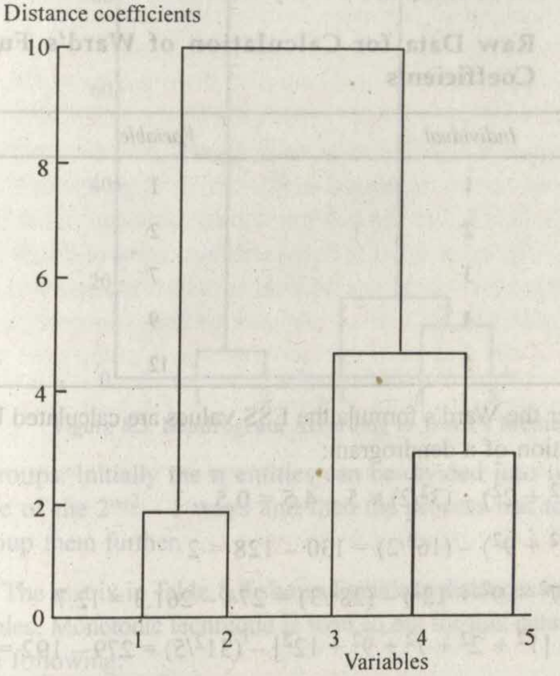


Fig. 8.2. Complete Linkage Dendrogram

Ward's Method of Clustering

J.H. Ward devised a method in which hierarchical grouping is done to optimise an objective function. First the variables are measured by the total sum of squared deviations of each point from the mean of the cluster to which it belongs, and then two clusters whose fusion values are in the minimum increase in the error sum of squares are combined.

Ward's formula for Error Sum of Square (ESS) is

$$ESS = \sum_{i=1}^{n} x_i^2 - \frac{(X_j)^2}{n}$$

An example is worked out for the Ward's method of computation of fusion values as per the data presented in Table 8.7. In this method one need not calculate the dissimilarity coefficients, the fusion coefficients can be calculated directly from the variables.

Table 8.7: Raw Data for Calculation of Ward's Fusion Coefficients

Individual	*Variable*
1	1
2	2
3	7
4	9
5	12

As per the Ward's formula the ESS values are calculated below for preparation of a dendrogram:

$ESS_{12} = (1^2 + 2^2) - (3^2/2) = 5 - 4.5 = 0.5$

$ESS_{34} = (7^2 + 9^2) - (16^2/2) = 130 - 128 = 2$

$ESS_{345} = (7^2 + 9^2 + 12^2) - (28^2/3) = 274 - 261.3 = 12.7$

$ESS_{12345} = [1^2 + 2^2 + 7^2 + 9^2 + 12^2] - (31^2/5) = 279 - 192 = 86.8$

On the basis of the above ESS values a dendrogram is prepared and shown in Figure 8.3.

Divisive Method

Divisive method is another simple method of clustering entities. There are two divisive techniques (a) Monotonic and (b) Polithetic. In the first technique clustering is made on the basis of possession or otherwise of a single specified attribute. In the second case the clustering is based on the value taken by all the attributes. Monotonic technique is easier and most popular with the cluster analysis.

Monotonic Division: In this technique first all the variables are divided into two groups and then the variables in the main group are examined one after another to find out their position in one of

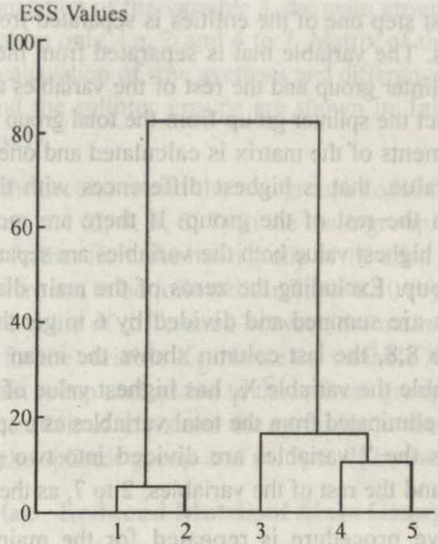


Figure 8.3. Dendrogram According to Ward's Method

the groups. Initially the n entities can be divided into two subsets, in one of the $2^{n-2} - 1$ ways and then the process has to be started to group them further.

The matrix in Table 8.8 shows Euclidean distances among seven variables. Monotonic technique is worked out for this data and shown in the following:

Table 8.8: A Matrix of Dissimilarity Coefficients

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	*Row Average*
x_2	0	10	7	30	29	38	42	26.00
x_2	10	0	7	23	25	34	36	22.50
x_3	7	7	0	21	22	31	36	20.67
x_4	30	23	21	0	7	10	13	13.33
x_5	29	25	22	7	0	11	17	18.50
x_6	38	34	31	10	11	0	9	22.17
x_7	42	36	36	13	17	9	0	19.50

In the first step one of the entities is separated from the group of total entities. The variable that is separated from the total group is termed as splinter group and the rest of the variables as the 'main' group. To extract the splinter group from the total group the averages of the row elements of the matrix is calculated and one of the rows with highest value, that is highest differences with the others, is separated from the rest of the group. If there are more than one row with same highest value both the variables are separated to form the splinter group. Excluding the zeros of the main diagonals each row's elements are summed and divided by 6 to get the arithmatic mean. In Table 8.8, the last column shows the mean value of the rows. In the Table the variable X_1 has highest value of 26.00 hence this variable is eliminated from the total variables as a splinter group. In other words the 7 variables are divided into two groups: 1 as splinter group and the rest of the variables, 2 to 7, as the main group.

The above procedure is repeated for the main group, i.e. calculation of row mean and elimination of the row with highest positive difference between main group and the splinter group (Main group average minus splinter group average) Table 8.9 shows the work table of elimination of next row.

Table 8.9: Separation of Second Variables

Variable	*Splinter Group (1)*	*Main Group (2 to 7)*	*Col.3—Col. 2*
1	2	3	4
2	10	25.00	15.00
3	7	23.40	16.40
4	30	14.80	−15.20
5	29	16.40	−12.60
6	38	19.00	−19.00
7	42	22.20	−19.80

In Table 8.9, the highest average difference between the main and splinter groups is 16.40 of row 3, hence the variable X_3 is separated from the main group and merged in the splinter group.

After the separation of the variable 3, the main group is now reduced to five variables only, i.e. 2 and 4 to 7. Matrix of this reduced main group with calculation of row averages and difference values between the main and the splinter groups are shown in Table 8.10 (a) and (b) respectively.

After the second round the two groups consists of (1, 3) in the splinter group and (2, 4, 5, 6, 7) in the main group. For elimination of variables further from the main group in the third round, again the same procedure has to be followed. In Table 8.10 (b) we find that the only positive value of difference between the main and the splinter group is 21 for the variable X_2. Hence the variable X_2 is seggregated from the main group and added to the splinter group. So at the third round of operation of divisions the variables in the two groups are (1, 2, 3) in the splinter group and (4, 5, 6, 7) in the main group.

Table 8.10(a): Reduced Matrix of Main Group

	X_2	X_4	X_5	X_6	X_7	*Average*
X_2	0	23	25	34	36	29.5
X_4	23	0	7	10	13	13.2
X_5	25	7	0	11	17	15.0
X_6	34	10	11	0	9	16.0
X_7	37	13	17	9	0	18.7

Table 8.10(b): The Third Round of Division

Variable	*Splinter Group Average (1, 3)*	*Main Group Average (2, 4, 5, 6, 7)*	*Difference between Main and Splinter Groups*
2	8.50	29.5	21
4	25.5	13.2	-12.3
5	25.5	15.0	-10.5
6	34.5	16.0	-18.5
7	39.0	18.7	-20.3

Now the main group with four variables of (4, 5, 6, 7) can be further splited, on the same procedures. But as it can be seen from Table 8.11, that all the difference values between the main and the splinter group are negative, which shows that there is no possibility of further division of the main group. Hence of the 7 variables the two clusters are (1, 2, 3) in the first cluster and (4, 5, 6, 7) in the second cluster. This is the final clustering of the variables. Within each cluster the variables can be sub-grouped further. In Table 8.11, we find that even though the difference values are all negative the variables (4, 5) have less difference values than the variables (6, 7) hence they can be separated further as sub-groups. In the first cluster also the average difference in distances are 26 for (1), 21 for (2) and 16.4 for (3), hence these three variables can be subdivided further into two groups as (12) in one and (3) in the other. The groups and sub-groups formed out of the 7 variables are shown in the link diagram in Figure 8.4.

Table 8.11: The Final Round of Analysis

Variable	*Splinter Group (1, 2, 3)*	*Main Group (4, 5, 6)*	*Difference*
4	24.6	10	−14.6
5	25.3	11.7	−13.6
6	34.3	10.0	−24.3
7	38.0	13.0	−25.0

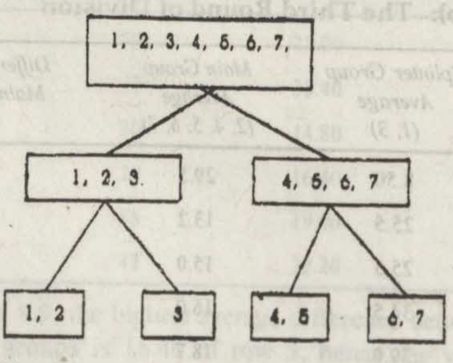


Figure 8.4. Link Diagram of Clusters According to Divisive Method

Selection of Clusters

Selection of an exact number of clusters which are significant for the study is the final task of a cluster analyst. A cluster analyst has to select only few clusters, which are significant, out of large number of clusters, for the final study. Specific principles for the selection of clusters have not yet been developed. Thorndike in 1953 adopted a graphical method for the selection of clusters. His method is based on fusion coefficients. This can be explained with the help of a diagram.

In Figure 8.5, along with the y-axis clusters are measured while in the x – axis mean of the fusion values of the clusters are measured. Fusion coefficient is a mean numerical value of the cluster's fusion values. The number of significant clusters are determined where from the cluster curve CC′ flattens, and almost parallel to the x axis. This selection test is similar to Scree test earlier described for the determination of factors in factor analysis. In Figure 8.5 the curve flattens from the point F, hence 3 clusters are regarded as significant for the study and rest of the clusters are eliminated from the study.

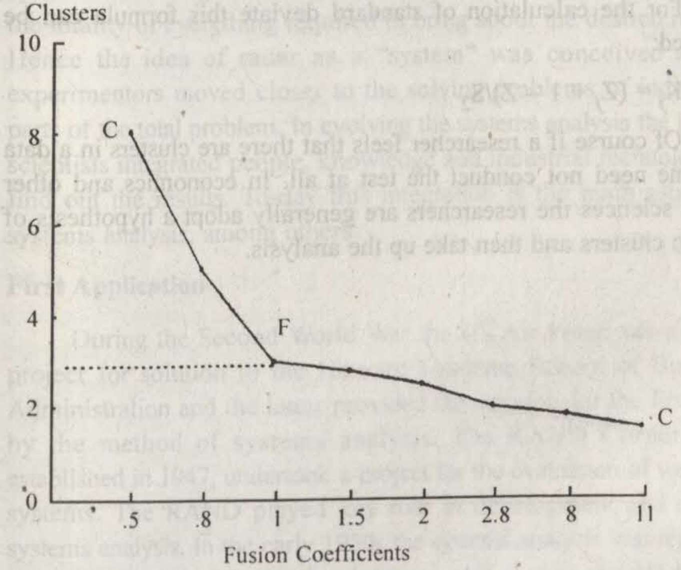


Figure 8.5 Graphical Method for Selection of Clusters

Cluster Tests

Before taking up any set of data for the cluster analysis one has to determine whether the data set has any clusters or not. If there are no clusters in the data set the researcher easily abandoned the data without any cluster analysis. This saves time of the researcher as well as energy because there is no point in taking a data set for cluster analysis if there are no clusters in the data set.

In 1977, R. Mojena familiarised a formula through which one can test whether there are clusters or not in a data set. According to the formula if the inequality prescribed cannot be satisfied the data have only one cluster.

Mojena's Formula $(Z_j + 1) > (Z + KS_2)$

Where Z is the mean value of the fusion coefficients:

$Z_j + 1$: Value of coefficient at stage j + 1.
K : Standard Deviate.
S_2 : Standard deviation of the fusion coefficients.

For the calculation of standard deviate this formula can be adopted:

$$\dot{K}_j = (Z_j + 1 - Z)/S_2$$

Of course if a researcher feels that there are clusters in a data set, one need not conduct the test at all. In economics and other social sciences the researchers are generally adopt a hypothesis of certain clusters and then take up the analysis.

9

Systems Analysis

The second world war, apart from its destructive character, initiated many new developments in science and technology as well as in research methodology. One such analytical tool developed during the war was systems analysis. The British scientists' original work during World War II in the field of radar technology laid the foundation of the modern systems analysis. In order to save England's cities from the enemy attacks the scientists made experimental and practical efforts, and as a result there emerged an interesting concept: the system. In its original form, the new concept of system contained the totality of everything required to bring about the desired results. Hence the idea of radar as a "system" was conceived as the experimentors moved closer to the solving problems of individual parts of the total problem. In evolving the systems analysis the British scientists integrated people, knowledge and industrial technology to find out the results. To-day this integration is the main aspect of systems analysis, among others.

First Application

During the Second World War the US Air Force submitted a project for solution to the Harvard Graduate School of Business Administration and the latter provided the solution for the first time by the method of systems analysis. The RAND Corporation, established in 1947, undertook a project for the evaluation of weapons systems. The RAND played key role in development and use of systems analysis. In the early 1950s the systems analysis was regarded as an inevitable tool for the defence engineers in the USA. The development of strategic missile systems and systems designed to detect and repell air attacks made great influence on the formulation of the technique of systems analysis.

The systems analysis became a popular method for the construction and production engineers. The integration of hardware components into complex electronic or electro-mechanical end products came to be known as systems engineering. This method has been called in various terms as systems design, systems analysis or systems approach. By 1960 the systems analysis emerged as a critical means by which total system is conceived, planned, manufactured and made operational. It encompasses all the functions that must be done and all of the operations; i.e. how the functions are to be implemented necessary to the total systems. Command and control systems become the important parts of the systems analysis. Command implies the capability of transmitting data to personnel and equipment through an information processing system. Control is the main objective of the systems functions and operations and implies the ability to execute decisions in a prescribed manner within a given environment. At the beginning, the military establishments institutionalised the systems analysis and raised it to the level of science using command and controls as the philosophy of systems design.

Use in Social Sciences

Use of systems analysis for problem solving in civilian life emerged in the middle of the 1950s. Development of governmental policy with regards to water resources by Ronald N. Mekean of the U.S. is one of the first attempts of systems analysis outside the defence departmental institutions. The systems analysis is now widely applied in economics, planning, management, business government policy, sociology and political science; to solve problems in their specific areas. The development of technically advanced hardware and systems projects are very much influenced by the application of operations research and econometric models. However the methods of formulating and applying systems analysis are not yet standardised. But the general need of the systems analysis is now well established.

What is a System?

The term 'system' is derived from Greek word 'Systema' means a "whole made up of parts". This term has different meanings in different subjects. But system as a method of analysis is known as systems analysis. Any aggregate of objects united in some form of regular interaction by nature or as to form an integrated whole to function, operate or move in unison and often in obedience to some form of control can be defined as systems analysis. In other words a set of components that work together for the overall objective of the whole is known as a system. Systems analysis is then a way of thinking and approaching problems that looking at the whole problem rather than concentrating on one or more parts to the exclusion of everything else.

Main Characteristics

The main characteristics features of a system are briefly mentioned in the following.

1. Systems are inter-related elements of work processes. Integration is the main aim of the systems. These elements are called sub-systems. One sub-system serves one or more higher order sub-systems and in turn be served by one or more lower order sub-systems.
2. The systems are mainly deterministic in nature. There are nothing haphazard or accidental in the integration of the sub-systems. The elements of a system fit together in a specific relationship and provide information in achieving an objective.
3. The inter-relationship of the sub-systems may be vertical or horizontal. The horizontal integration takes place across the sub-systems, while the vertical integration involves the lower order sub-systems and sub-systems to system.
4. The sub-systems in a system operate in a particular order and remain responsible for the function of the system. The influence of the whole on the parts or the sub-systems is also there in a specific proportions.

5. ***Stability:*** Some systems are stable in nature. After a shock to a system when it comes to its original state again it is known as a stable system, otherwise it is known as unstable system.
6. ***Feedback:*** Some systems have feedback mechanism. Some of the output of the system goes back and influences the systems output in the form of inputs. If there are several feedbacks in a system it becomes more complex for control.
7. Some systems are self-regulated. If any of the sub-systems go out of the way and it is dragged back by other sub-systems or the system itself functions in such a manner that automatically regulates on its own it is called a self-regulated system. Most of the natural systems are self-regulated systems up to a certain level.

Classification of Systems

Systems can be classified on the basis of their nature and functioning. Some of the important classifications are given below:

1. Systems can be classified broadly as nature-made and man-made systems. The nature-made systems are those which operate on some natural phenomena on which man has no power to influence its primary functions or regulate the system. The solar system in which different planets orbit the Sun in a fixed proportion of time and speed, is a nature-made system. The biological system of human beings or animals are also nature-made on which man cannot influence its operations of any fundamental nature.

   Economic, political, social and administrative systems are completely man-made. Accordingly man can regulate its structure and function and can make fundamental changes required for specific needs of the society.

2. Systems can be divided into two type again as physical systems and conceptual systems. A dynam which generates power is a man-made physical system. It can be regulated and controlled by men even though it is machine dominated. There are also man-machine systems. The fighter aircraft whose functions are

integrated with the pilots is a man-machine system. One sub-system of fighter aircraft is a range computer, which is associated with a radar sub-system into which the pilot, another sub-system, makes the initial inputs. The car driver and his car is a man-machine system.

An economic system is a conceptual system. A model for economic development of a country with given structure of the society and economic institutions is a conceptual system. The conceptual systems are man-made and probabilistic in nature, since it involves a social structure with which end results can be predicted accurately.

3. Systems can be classified on the basis of their openness. If the influence of the environment on the system is assured and systems analysis is done, it is known as open system. On the other hand, if it is assumed that the system has no influence on the environment or vice versa, it is regarded as a closed system. In economics when a country's economy is analysed without taking into account its interaction with the rest of the world it is a closed system. Open systems are more complex than the closed ones.

Systems Methodology

A system is determined by a given set of objects, properties and their relationships. The main systems objectives are: input, output, process, feedback and restriction or control.

Inputs: Every system has an input-output mechanism. Input is defined as the energizing or the start up component on which the system operates. In physical systems such as motor, runs on the basis of oil to mobilise the machine. In economic systems investment, income, and employment created in the economy are the inputs to generate a specific degree of standard of living and money incomes in the economy.

Outputs: Output is deffned as the ultimate result of an operation. In other words output stands for the purpose objective or goal of the system for which a system is operated. A system is operated only for the outputs. Data fed to a computer with regard to the income of the families in a locality is input, and if we want the average income pattern of the locality the data will be processed in the computer and the result, the average income, comes out of the computer is known as output. The output depends upon the volume and nature of the input and also on the structure and capacity of the system.

Process: The process is defined as the activity that makes possible the transformation of input into output. It is a phenomenon that manifests continuous change over time. In a production system, for instance, the raw materials and the labour are the inputs, the plant and machinery which processes them are the processors or hardware, and the finished products are the output. To know the structure and function of the processor is an important aspect of systems analysis. In some natural systems, functions are not known but we know that it transforms a specific input into output. When a system is not known except its input-output behaviour the system is known as 'black box'.

Feedback: Feedback in a broader meaning is output redelivered as input and introduced in the system from which the output is derived. Output of a sub-system reintroduced into the system or in another sub-system, are also known as feedbacks. A feedback performs a number of operations:

(a) It compares the actual output with assumed output.

(b) It identifies the discrepancies that exists between the two.

(c) It evaluates the content and significance of these differences.

(d) It computes the required input for use as new systems input.

When some of the finished products of an industry are used again in the same industry as input it is called feedback input. The main objective of feedback is to control the system; to modify the inputs with the objective of improving the systems performance and output. There may be a single feedback or multiple feedback as per the structure of the system.

Input, output, process and feedback mechanism are shown in Figure 9.1. There are three feedbacks in the system shown in the diagram. First is from D to A, second from C to A and third from B to A. This is a case of multiple feedbacks.

Restrictions: Systems function under certain restrictions and limitations. The restrictions to a system stipulates internally or externally imposed guidelines that bound the problem. The restriction is composed of two elements: the objective and constraints. The objective is defined as the goal or output to be achieved. The constraint is the factor that limit the objective. In a macrolevel production system, the objective is to reach a particular output, while availability of capital and manpower are the constraints. The constraints must be consistent with the objectives and not mutually exclusive.

A system operates by its sub-systems. All the sub-systems necessary to a system describes the system's boundary. The boundary of a system identifies its scope, and also determines the totality of inputs, process and outputs required to operate a given system.

Regulation and Control: The main purpose of systems analysis is to control or regulate the system. When an influence is exerted on the results of operations of the system in order to achieve a desired result or output, it is known as control of the system. The term regulation is often used in a narrower sense of steering or directing the system to get certain results. One of the main mechanisms for regulation is the feedback paths in the system. The feedbacks regulate the system, hence by controlling the feedbacks one can control the system.

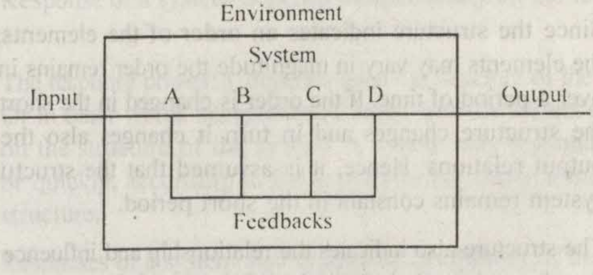


Fig. 9.1. Feedbacks in the system

Systems Theory

Systems theory is an intellectual tool for studying the relationship between the structure of a system and its functioning. Since the main objective of systems analysis is to control the system for the desired output, the first task of a systems analyst is to identify the functioning of a system with a known structure. If the structure of a system is not known the analyst has to specify a possible structure for the system.

The three basic concepts of a system to be analysed are:

1. What is the composition of the structure?
2. How the structure functions?
3. How it responds to the inputs?

Structure of a System

A system is composed of certain sub-systems which are called elements. A structure can be defined as an ordered set of inter-connected operations performed by the elements of a system. A model builder in fact prepares a structure on the observed pattern of activities of the elements. The three important characteristics of a structure are:

1. A structure characterises a specific arrangement or an order of parts or elements rather than their concrete nature. Inputs introduced in the system move in a given path towards output through the functions of the elements and their inter-relationships. The elements may be in series or in parallel make specific operations to influence the system to provide output.

2. Since the structure indicates an order of the elements, while the elements may vary in magnitude the order remains invariant over a period of time. If the order is changed in the short period the structure changes and in turn it changes also the input-output relations. Hence, it is assumed that the structure of a system remains constant in the short period.

3. The structure also indicates the relationship and influence among the elements and also certain rules on which the various parts can be combined in the system. Any group or combination of elements cannot make a structure, unless the elements have certain specific relationship and their combination observes certain rules.

Function of a System

Function denotes a relation between a part of a system and some selected 'state' of the whole. A function is then a mode of transformation of inputs into outputs; or the rules by which the elements of the set of inputs are associated with the elements of the set of outputs.

Function of a system indicates two information:

1. It indicates the rules according to which the elements of a set are associated with other elements. Systems transform variable inputs into variable states, and the variable states into variable outputs by the invariant transformation rules.
2. Function explains also the mode of transformation. The transformation is the main activity of a system. The systems output is realised on the nature and basis of transformation.

Response of a System

Response of a system is its output. The system yields an output given a particular history of inputs. It is always a response at a particular moment of time. In other words change over time is the central force of systems analysis. Systems response to the variations in the external environment, enters the system as input. Different systems respond differently to the inputs.

Response of a system depends independently on the following factors:

1. The response primarily depends upon the structure of the system. Or in other words acceptance of inputs into the system depends on the structure of the system. A system may respond slowly or quickly, according to the activity of the sub-systems in the structure.
2. Responses of a system also depends on the history of the inputs. The manner the inputs are introduced into the system determines the degree of response.

The system analysis starts with the model of the structure or those aspects of systems that are synchronic; with the knowledge of synchrony it studies the dynamic changes of the states of a system; and finally the responses to inputs.

Input-output States

In systems analysis U(t) is denoted as systems' input at a given point of time which is called a 'state' that is, 'input state'. And Y(t) is denoted as systems' output at a given point of time which is called 'output state'. Of course the latter's time is different from that of the input time since the systems take time to transform input states into output states, that is, a system works over a time.

On the basis of number of input and output states, systems can be divided into two: Simple or Single state system and Multiple state system. When there is one input state and one output state it is called single state system, on the other hand if there are multiple inputs and multiple outputs it is known as multiple system.

Intermediate Outputs: When a system is a complex one it produces some intermediate outputs; and these outputs become intermediate inputs within the system. These internal inputs and the outputs are also 'states' of the system. Internal states of a system are the partial results of transforming the inputs as intermediate outputs.

State Variables: In order to analyse the dynamic behaviour of a given system one has to analyse the state variables. It is required because they produce information required to predict the future behaviour of the system on the basis of the past variables.

Elements of a System

In Figure 9.2, the O_1, O_2 are the operations and the X_1, X_2 are the state variables or elements. The two important characteristics of operations are:

1. The operations are ordered
2. And they are interconnected.

The connections are shown in solid lines which are called paths. The paths are divided into two, on the basis of their direction: as

(a) Forward Path and (b) Feedback Path. When the operations are ordered from left to right that is, from inputs towards outputs it is known as forward path. Generally arrow marks are not drawn in the forward path as it is assumed that the direction is from left to right. Operations right to left, or the reverse path, that is from outputs towards inputs is known as backward path. An arrow mark direction is invariably given in the backward path. These operations convert one state variable X_i to another state variable.

The forward and backward paths are shown in Figure 9.2. The operations O_1, O_2 and O_3 are in the forward path, while the O_4 is in the backward or feedback path.

Interconnections

The interconnections in the paths may be divided into two categories: (a) single connections with or without feedback path and (b) multiple connections with or without feedback path. It is these interconnections that determines the structure of a system. The ordered sets of interconnected operations performed on the forward path is termed as boundary structure and the interconnected operations performed on the feedback paths is known as characteristic structure. In the Figure 9.2, the system is a characteristic structure since there is one feedback path. Of course in a system all the paths ultimately have to touch the output as a forward path. The feedback paths act as controlling mechanism. A system may or may not have a feedback path but there cannot be any system without a boundary structure.

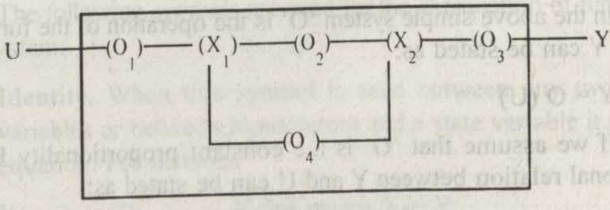


Fig. 9.2. Operations and Paths

Function of a System

The function of a system as stated earlier is the mode of transformation of inputs into outputs, resulting from the total combination of operations applied by the system. Hence the function of a system can be thought as "rules of transformation" that given the objects in the domain of inputs associated with them an unique object from the range of outputs. Systems transform inputs into outputs in a specific manner or pattern; this manner of transformation is known as function of a system.

The two important characteristics of function of a system are:

1. In the short period, since the technology remains constant, the function of a system is assumed as invariant, but the inputs and the outputs can vary as per the states of the inputs.
2. With a given function the ordering of outputs is completely determined by the ordering of the inputs over time. As the technology remain constant in a short period the ordering of the state variables also remain constant, or in other words the interconnections among the sub-systems remain constant.

Rules of Operations

Two simple rules of operation in a system are the rules of multiplication and addition. In the input-output relation the 'O' that is, operation is the function of a system.

Rule of multiplication: Rule of multiplication says that one has to multiply the operations along the same forward path to determine the function of the system.

U____________________(O)____________________Y

In the above simple system 'O' is the operation of the function; hence Y can be stated as:

$Y = O\,(U)$

If we assume that 'O' is the constant proportionality K, the functional relation between Y and U can be stated as:

$Y = K\,U$

In the following system there are two functions along the same path:

U __________ (O_1) _________ X_1 _ _ _ _ _ (O_2) __________ Y

In this system the first function O_1 transforms U into state variable X_1 and the second function O_2 transforms the state variable X_1 to output Y. Hence the two functions together can be stated as follows:

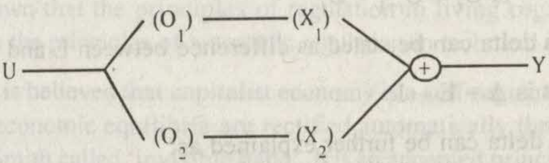
$X_1 = O_1U$ and $Y = O_2 X_1$

$Y = O_2 (O_1U) = (O_1 O_2) U$

Hence the function of this system is multiplication of first operation O_1 with the second operation O_2. If function is denoted as F;

$F = O_1 O_2$ so $Y = FU$, this is the rule of multiplication.

Rule of Summation: When the operations are parallel to each other or when the lines come together before going to output states the operations are added to make the function of the system.



From the previous example of multiplication rule we know that $X_1 = O_1U$ and $X_2 = O_2U$ hence:

$Y = X_1 + X_2 = O_1U + O_2U = (O_1 + O_2)U$

Here function $F = O_1 + O_2$ or $Y = FU$

The following symbols are used for the explanation of functions in a system:

I : **Identity.** When this symbol is used between any two state variables or between input/output and a state variable it means equation. For instance

$X_1 \underline{\qquad}(I)\underline{\qquad} Y$ this means $X_1 = Y$.

K : **Proportional transformation.** Between any two state variables this symbol indicates the transformation of one state variable to another with a constant proportion whether in the forward path or the backward path.

E : **This indicates forward difference or advancement in time only.** This is also known as lead operator or shift operator. If a collection of the possible values of the input of the system can be arranged in the form of a sequence as $U_1, U_2, U_3 \ldots U_n$ then the operator E transforms the state of input U_i into the state of output U_{i+1} and therefore $EU_i = U_{i+1}$. The backing operator or the inverse shift operator is analogous to E, but here the state of input U_i is transformed into state of output U_{i-1} and this is denoted as E^{-1}; therefore $E^{-1}U_i = U_{i-1}$.

Δ : **The differentiating operator or the finite difference operator.** The meaning of this operator is that. if a collection of the possible values of the input of the system is presented as the sequence $U_1, U_2, \ldots U_n$ then the operator Δ transforms state of input U_i into the difference $U_{i+1} - U_1$ this means:

$$\Delta U_i = U_{i+1} - U_i$$

This delta can be stated as difference between E and 1.

That is $\Delta = E - 1$

The delta can be further explained as:

E	=	1	=	Δ
$U(t+1)$		$U(t)$		$U(t+1) - U(t)$
$a + b(x+1)$		$a + bx$		b
$x + 1$		x		1
c		d		$c - d$

Since Δ is the differentiating operator $\Delta - 1$ can be accepted as accumulation. That is

$$\Delta^{-1} = \frac{1}{E-1}$$

Therefore $Y(t) = 1/(E-1)\,U(t) = \Delta^{-1}\,U(t)$

D : **The differential operator.** Assuming that the state of input U is a function of certain parameter t that is, $U = f(t)$, this operator means that in order to obtain the state of output we should differentiate the function expressing the state of input, i.e. determine its derivative. Hence the differential operator is denoted by the symbol d/dt or simply by D. According to differential calculus the determination of derivative satisfies the condition of linearity of transformation, because DcU=cDU.

Regulation and Control

The main purpose of systems analysis is to regulate and control the system. The science of control or cybernetics is a new science originated in 1948 with the publication of the book *"Cybernetics or Control and Communication in the Animal and the Machine"* by Norbert Wiener. He was a famous mathematician, the founder of cybernetics, discovered the existence of profound analogies, between the functioning of servo-mechanism (or automatic control) and electronic computers and functioning of living organisms. He has shown that the principles of regulation in living organisms are same as the principles of automatic regulation in technical equipment.

It is believed that capitalist economy is a self-regulated system, where economic equilibria are rectified automatically through what Adam Smith called 'invisible hand'. It is an accepted principle to-day that the technique of automatic regulation and control which is applied to living organisms and electronic machines, can be successfully introduced in socio-economic processes. Thus, cybernetics came into being as a general science of regulation and control of the systems consisting of various mutually related elements or sub-systems.

In cybernetics the term 'control' is used in the meaning that some influence is exerted on the results of operations of the system in order to achieve a desired output. On the other hand the term 'regulation' is often used in a narrower sense of steering or directing the system for control.

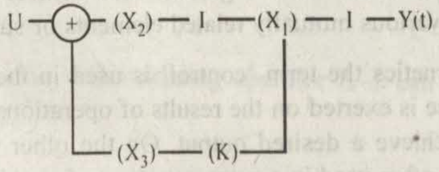
A system can be regulated by three of the following methods:

1. Regulation can be performed by equalization of deviations of the real values of output from the required value of output. This principle is called the equalization of deviations. Devices used for this kind of regulation are called Regulators.

2. Regulation can be performed by the compensation of disturbances. When the environment influences the system, this disturbance can be regulated by making necessary compensation or compensating the influences by devising certain regulating mechanism inside the system. When heavy outside cold affects the body, it is compensated by taking necessary medicines and liquids that generates heat inside the body. Devices for compensating disturbances are called compensators.
3. Regulation can also be performed by the elimination of disturbances. In this regulation the effects of environment is eliminated by protecting the system from the attacks of external elements. These devices are called by various names such as shock observers, buffers, shields and insulators etc. When we put on warm clothes to protect our body from cold this can be an example for elimination of disturbances. When an economy is protected from external competition through the restrictions on imports this is also elimination of disturbances in the economic system.

Regulation based on equalisation of deviations is characterised by feedback. It is the feedback which regulates the system.

A simple control system mechanism is explained here in the following regulated system:



Function of this system can be derived as under:

$X_2 = U + X_3$; $X_1 = X_2$, $X_3 = KX_1$; $Y(_t) = X_1$

Accordingly, $X_2 = U + KX_1$ (Since $X_3 = KX_1$)

$= U + KY$ (Since $X_1 = Y$)

$Y = U + KY$ (Since $Y = X_1 = X_2$)

$$Y - KY = U$$

$$Y\,(1\ K) = U$$

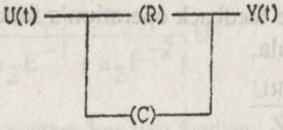
$$Y = \frac{1}{1-K}U\ ,$$

Therefore the function $F = \dfrac{1}{1-K}$

In the feedback system K is the proportional transformer. In this function if K = 0, that means U = Y; and the magnitude of input and the output is the same, but if K = 1, the output will be half of input. Of course when K is more than one the output will further reduced proportionally. Hence the K is the regulator of output.

Mason's Formula

Functions for any type of structure in a system can be easily found on the basis of Mason's rule. This can be explained in a simple system's structure. R is denoted as forward path operation and C as the backward path operation. Accordingly a formula can be derived from the diagram.



The output equation is:

$$Y = R\,(U + CY)$$

$$= RU + RCY$$

$$Y - RCY = RU$$

$$Y(1 - RC) = RU$$

$$Y = \frac{RU}{1-RC} = \frac{R}{1-RC}U$$

Hence this formula can be used for any single feedback operation and is called Mason's rule:

$$F = \frac{\text{Forward path operation}}{1 - (\text{Forward path operation})\,(\text{Feedback path operation})}$$

In the formula there are two parts: $\frac{1}{1-RC}$ determines the operation of the feedback path which is called feedback multiplier and the second factor RU determines the operation of the main forward path. Whatever may be the operation in the R and C, out of the four operations mentioned earlier one or all of them together, can be used in the formula. Hence this is regarded as the basic formula of the theory of regulation.

In the system the state of actual output can be known for a given state of inputs. This formula can be used also for realisation of a desired output Z. Instead of Y it can be put to Z. In order to realise Z what magnitude of U is required can be calculated on the above formula.

$$Z = \frac{R}{1-RC} U$$

Hence $U = \frac{1-RC}{R} Z$

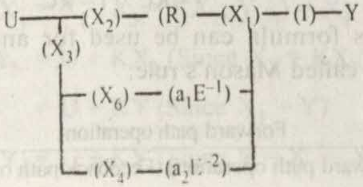
This shows that when the reverse of the function for the feedback path and main forward path with multiplication of desired output the magnitude of input state can be known. Similarly from the above formula the feedback operation's function C can be known from the basic formula.

Where $C = \frac{Z-RU}{RZ}$

This shows that for a given values of Z, U and for the specific operation of R, one can ascertain the nature of feedback operation.

Mason's Rule can be utilised for any type of structures for the calculation of the functions. Three of systems structures and their calculation of functions are analysed below:

System I: In this system there are two backward paths and they are parallel in nature.



$X_5 = a_1E^{-1}X_1$; $X_4 = a_2E^{-1}X_1$ and $X_3 = X_5 + X_4$

$X_2 = U + X_3$; $X_1 = R(\Delta)\, X_2$ and $Y = X_1$

Hence $X_3 = a_1E^{-1}X_1 + a_2E^{-2}X_1$

$$= X_1\,(aE^{-1} + aE^{-2})$$

$$= Y\,(a_1E^{-1} + a_2E^{-2}) \text{ since } X_1 = Y$$

This is the combined feedback path's function:

$a_1E^{-1} + a_2E^{-2} = C(\Delta)$

Next in the forward path:

$$Y = R(\Delta)\, X_2$$

$$= R(\Delta)\,(U + X_3)$$

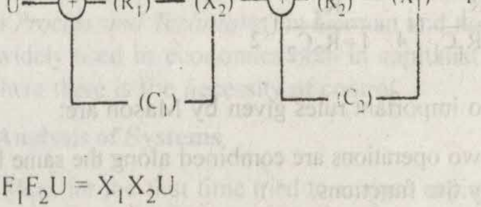
$$= R(\Delta)\, U + R(\Delta)\,(a_1E^{-1} + a_2E^{-2})\, Y$$

$$Y - R(\Delta)(a_1E^{-1} + a_2E^{-2})\, Y = R(\Delta)\, U$$

$$Y\,(1 - R(\Delta)(a_1E^{-1} - a_2E^{-2}) = R(\Delta)\, U$$

$$Y = \frac{R(\Delta)}{1 - R(\Delta)(a_2E^{-1} + a_2E^{-2})}\, U$$

System 2: This system has two feedback paths but in a different structural position than the feedbacks in the earlier system. But same Mason's formula can be used to get the function. The system structure is in the following way.



$$Y = FU = F_1F_2U = X_1X_2U$$

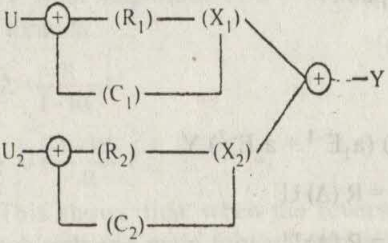
$$X_2 = F = \frac{R_1}{1 - R_1C_1};\; X_1 = F_2 = \frac{R_2}{1 - R_2C_2}$$

As per rule of multiplication:

$$F_1F_2 = \frac{R_1R_2}{(1-R_1C_1)(1-R_2C_2)}$$

$$\text{Hence } Y = \frac{R_1R_2}{(1-R_1C_1)(1-R_2C_2)}U$$

System 3: In this system there are two feedback paths but the structure is different. The Mason's rule can be applied here where the forward paths are in parallel.



$$X_1 + X_2 = Y$$

$$X_1 = \frac{R_1}{1-R_1C_1}U_1$$

$$X^2 = \frac{R_2}{1-R_2C_2}U_2$$

Hence the function of this system is:

$$Y = \frac{R_1}{1-R_1C_1}U_4 + \frac{R_2}{1-R_2C_2}U_2$$

The two important rules given by Mason are:

1. When two operations are combined along the same line of flow multiply the functions.
2. When two operations are combined by two different lines the functions are added together.

On the basis of these, functions of any complex structure of system can be analysed without any difficulty.

Systems Analysis in Economics

Earlier economists, even though not aware of the science called cybernetics or systems analysis used the mechanism of regulation and control in economic analysis. Adam Smith, the first economist to study the mechanism of an economy, believed that the economy is self-regulated, and hence there is no need of control from outside. His regulatory mechanism was "laissez faire" policy and that he believed as the "invisible hand" that automatically controls the imbalances in the economy. According to him the state should not interfere in the course of economic process because it would disturb the mechanism of self-regulation and upset the automatically achieved economic equilibrium.

In mid-nineteenth century Karl Marx rejected the concept of self-regulation. He proved that the automatic mechanism of self-regulation of a capitalist economy is only a link in the dialectic process of development which leads to increasing internal contradictions in the economy. Even if self regulation operates it does not necessarily lead to socially desirable ends. Hence state intervention is necessary for controlling the pattern of economic process in a desired way.

Cybernetics as science of controlling and regulating system was established by a prominent mathematician Norbert Wiener. Tustin, an electrical engineering professor, has first shown how the theory of electrical network can be applied to the studies of economic systems and controlling and regulating those systems. But the main break came in 1957 with the publication of collected essays as "*Economic Regulation Process and Technique*" in German and now the control theory is widely used in economics both in capitalist and socialist systems where there is the necessity of control.

Marxian Analysis of Systems

Karl Marx for the first time tried to analyse capitalist economy through systems approach. He analysed the society in terms of economic activities based on certain interrelationship which determines its structure. Accordingly development of the society at different stages depend on certain specific economic systems. He divided the history of the society into five stages and each stage has a different economic system based on the social structure. The five stages described by him are: primitive communism, slavery, feudal economy, capitalist economy and communism. In the primitive communist system, there were no organised sub-systems, but in later stages of the development of the society the sub-systems become organised which in turn can be termed as systems. Marx in his *Das Capital* discussed elaborately the working of capitalist system.

Marxian analysis of "law of value" as a regulator of commodity production, in fact was dealing with the problem of regulation of the economic system and thus with cybernetics. His explanation of working of capitalist economy through the reproduction schemes is a good example of systems analysis. Oskar Lange made a detailed study of Marxian Reproduction Schemes and explained both the simple and expanded reproduction schemes in terms of control systems theory, and further developed it as dynamic theory of growth.

The production system defined by Marx in a capitalist economy is

$$X = c + v + s \qquad \ldots (1)$$

where the total product X expressed in value units is composed of three components; c constant capital, v outlay of direct labour and s the surplus value created thereof by labour. The coefficients of outlays can be introduced as under as,

$a_c = c/x$ and $a_{v \div s} = (c = v)/x$ and addition of both the terms is equal to one:

$$a_c = a_{v \div s} = 1$$

The equation (1) can be presented in the following form:

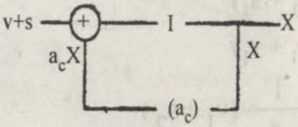
$$X = a_c X + (v + s)$$
$$X = a_c X = (v + s)$$
$$X\ (1 - a_c) = (v = s)$$

$$X = \frac{1}{1 - a_c}(v + s) \qquad \ldots (2)$$

This shows that there is a feedback relationship in the production process identical to systems formula of

$$Y = \frac{1}{1-RC} SU$$

The production scheme can be shown in terms of systems diagram in the following;



It follows from the figure above that direct labour v+s is transformed into product X and its identity transformation denoted by I. The regulated system is governed with the proportionality operator a_c whose existence is due to the fact that part of the product X must be used to replace the used up means of production.

The model can be extended to a two sector model of the national economy according to reproduction scheme used by Marx. The subscript 1 indicates the capital goods sector and 2 the consumer goods sector in the economy.

$$X_1 = c_1 + (v_1 + s_1) = a_{1c}X_1 + (v_1 + s_1) \qquad \ldots 3$$

$$X_2 = c_2 + (v_2 + s_2) = c_2 + a_2\,(v_1 + s_1)X_2 \qquad \ldots 4$$

The equilibrium condition of simple reproduction scheme is

$$c_2 = v_1 + s_1$$

This is because the demand and supply of capital goods are in equilibrium in the following condition:

$$c_1 + c_2 = c_1 + v_1 + s_1$$

$$c_2 = v_1 + s_1 \qquad \ldots(5)$$

The left hand term of the above equation is demand for capital goods and the right hand term the total supply of capital goods in the economy. This equilibrium is true also for the consumer goods sector. The equilibrium condition explains that the value of means of production which sector 2 acquires from sector 1, i.e. $v_1 + s_1$ must be equal to the value of the means of consumption transferred from sector 2 to sector 1, i.e. c_2.

From the equation (3) the systems operators can be derived in the following:

$$X_1 = a_{1c}X_1 + (v_1 + s_1)$$

$$X_1 - a_{1c}X_1 = (v_1 + s_1)$$

$$X_1 (1 - a_{1c}) = (v_1 + s_1)$$

$$X_1 = \frac{1}{1-a_{1c}}(v_1 + s_2)$$

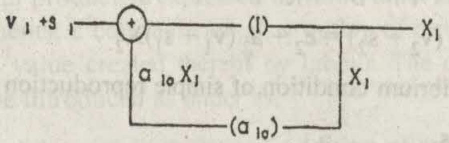
Similarly in the consumer goods sector the equation (4) is

$$X_2 = c_2 + a_2 (v + s) X_2$$

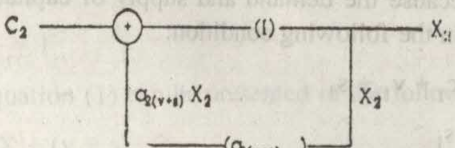
$$\frac{1}{1-a\,(v+s)}c_2$$

The equations (3) and (4) are presented in the form of block diagrams for sector 1 and sector 2 below:

Sector: I



Sector: II



On the basis of these equations the ratio of values of total products in both the sectors can be calculated. As it is a fact that $a_{2c} (v + s) = 1$, hence the production ratio between the two sectors in the economy can be calculated in the following:

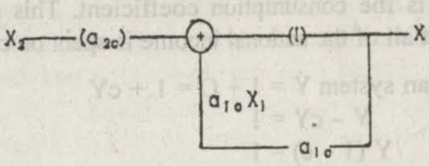
$$\frac{X_1}{X_2} = \frac{1-a_{2(v+s)}}{1-a_{1c}} = \frac{a_{2c}}{1-a_{1c}}$$

Hence

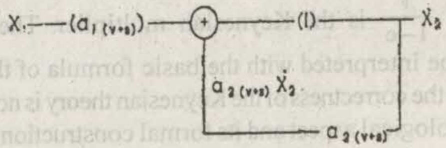
$$X_1 = \frac{a_{2c}}{1-a_{1c}}.X_2 \qquad ...(6)$$

$$X_2 = \frac{a_{1(v+s)}}{1-a_{2(v+s)}}.X_1 \qquad ...(7)$$

Sector: I



Sector: II



The equations (6) and (7) are presented in the form of block diagrams:

The block diagram can be easily interpreted from economics point of view in a capitalist economy. To produce X_2 of consumer goods it is necessary to have c_2 or $a_{2c}X_2$ capital goods where a_{2c} is the coefficient of outlay of capital goods in sector 2. This transfer takes place in a system with the operator a_{2c}. However to produce c_2 we need again certain amount of capital goods and therefore the system must be connected by feedback, in serial way, with the regulating system whose operator is a_{1c} i.e. the coefficient of outlay of the capital goods industry.

Similarly in the above figure the total production of X_1 requires an additional amount of consumer goods, to make the system in equilibrium. Therefore a regulatory system connected by a feedback is required. The feedback operator is $a_{2(v+s)}$ i.e. the coefficient of outlay of direct labour in sector 2.

The Keynesian System

Keynes in his "General Theory" explained a new interpretation of the working of the economy in terms of national income. Keynes considered the national income Y as a sum total of net expenditure in the national economy as sum of two components: Expenditure on investment 1, and expenditure for consumer goods C. This C can be treated as a linear function of national income, namely $C = cY$, in which c is the consumption coefficient. This satisfies the condition that not all of the national income is spent on consumption.

In Keynesian system $Y = 1 + C = 1 + cY$

$$Y - cY = 1$$

$$Y(1 - c) = 1$$

$$Y = \frac{1}{1 - c} 1$$

The term $\frac{1}{1-c}$ is the Keynesian multiplier. The Keynesian multiplier can be interpreted with the basic formula of the theory of regulation. Here the correctness of the Keynesian theory is not questioned only its methodological aspect and its formal construction is analysed.

According to Keynes, a system is influenced by certain investment activity l. This investment results in expenditure equal to the level of investment outlay. There is a feedback relation between the system with transmittance c. After the correction introduced by the governor the overall input of the first system is the activity of the size $Y = 1 + cY$. This is shown in the form of a block diagram in Figure 9.3.

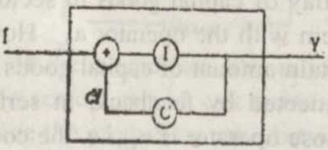


Fig. 9.3 The Keynesian Multiplier

If Y is to have a certain desired value z the input magnitude l amounts to the following formula, i.e. the systems operator or the Keynesian multiplier is reversed: l = (1 – c) Z.

10

Network Analysis

Network, connecting threads into knoghts in production of a net, used in analysing relationships between individuals, objects or events has been developed as a separate study called "network analysis". Any relationship arranged in the form of a network and its analysis is the main theme of network analysis. The theory of graph is the basis of network technique. It is widely used in social science research. In economics and other social sciences two of the most important applications of network analysis are: (1) Social behaviour analysis or sociogram and (2) Activity analysis. Economic activity is mainly based on the social behaviour or the behaviour of the individual, consumers and producers in the society. Analysing the behaviour of the individuals and the groups through the network of social relations is known as sociogram. The network analysis technique used in project planning and execution of work or activities over time is called activity analysis. The social behaviour analysis which uses network technique is discussed in this chapter, followed by activity analysis in the next chapter.

Principles of Graph Theory

The network drawings are based on graph theory. Hence some basic knowledge of graph theory is essential for layout of networks for analysis. A brief introduction of graph theory is described here.

A line joining two points is called a liner graph or simply a graph. The lines need not be a straight line, as supposed in linear alzebra, it is just a line in any form. A line has two ends or edges, hence in graph theory a line is denoted as an 'edge'. The edges connect two points or dots. These dots are known as 'vertex'. Hence a graph is defined as diagram consisting of edges and vertices. The G (Graph) is composed of V (Vertices) and E (Edges). Mathematically they are expressed as:

$G = (V, E)$

$V = (v_1, v_2, v_3 \ldots\ldots)$ vertices

$E = (e_1, e_2, e_3 \ldots\ldots)$ edges

An edge is identified as an unordered pair of vertices v_i, v_j. If an edge connects only the same vertex or a single vertex the edge is called a 'self-loop'. If two edges connect the same pair of vertices the edges are referred as parallel edges. In Figure 10.1 e_1 is loop while the edges e_4, e_5 are parallel edges. A graph which has no self loop or parallel edges is called a 'simple graph'. But if one of these two or both are present in a graph it is referred as a 'General Graph'. Accordingly the Figure 10.1 is a general graph.

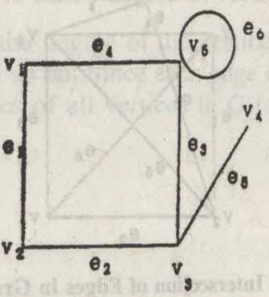


Fig. 10.1 General Graph

Characteristics of a Graph

1. A graph is identified by the number of edges and vertices it contains. The manner of lines or the length of lines connecting the vertices are immaterial. In Figure 10.2 all the four diagrams according to graph theory are same eventhough they are connected in different shapes, because they have equal number of edges and vertices.

2. In connecting edges with vertices sometimes the edges have to be intersected; in graph theory the intersection of the lines

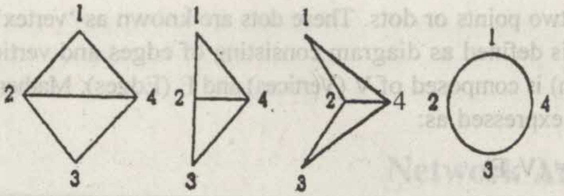


Fig. 10.2 Equal No. of Edges and Vertices

are not recognised, even if there is an intersection, both the edges are treated separately; there is no influence of one edge on the other. So any type of intersection among the edges are not accepted as intersection. In Figure 10.3 there are two intersecting edges, e_6 and e_7 with edges e_5 and e_2, but the graph has to be identified with number of edges and vertices only and not by its intersections. In the Figure 10.3 there are 5 vertices and 8 edges.

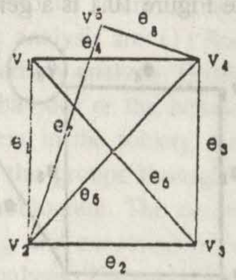


Fig. 10.3 Intersection of Edges in Graph Theory

Types of Graphs

Graphs can be divided into two classes: finite graphs and infinite graphs:

1. A graph with a finite number of edges and vertices is called a finite graph. This graph has specific number of edges and vertices. Unless otherwise mentioned a graph means always a finite graph.
2. A graph which has no finite number of edges or vertices is called an infinite graph.

3. **Null graph:** In a graph if there are only vertices but no edges it is called a null graph. Here E = ϕ. Of course there cannot be a graph without edges. Hence the null graph consisting of only vertices is also called vertex graph.

Some Concepts of Graph Theory

Incident: When a vertex is an end vertex of many edges the edges are called incident to each other. In the Figure 10.3 the edges e_2, e_6 and e_7, are incident with v_4.

Adjacent: Two non-parallel edges if they are incident on a common vertex, they are called adjacent edges. Similarly if two vertices are said to be adjacent if they are the end vertices of the same edge. In the Figure 10.3 e and e are adjacent, so also e_4, e_6 but e_4 and e_2 are not adjacent. The vertices v_2 and v_4 are adjacent while v_2 and v_3 are not adjacent.

Degree: The number of edges incident on a vertex with self loops counted twice, is called degree of vertex: $d(v_i)$.

In Figure 10.1 the degree of the vertices are $d(v_1) = 3$; $d(v_2) = 4$ and $d(v_4) = 3$ and so on. Since each edge contribute two degrees the sum of the degrees of all vertices in G is twice the number of edges in a graph.

$$\sum_{i=1}^{n} d(v_i) = 2e$$

Isolated Vertex: A vertex with no incident edge is called an isolated vertex. Even though it is a part of the graph, it is not connected to the graph with any other vertex with the edges. This is also known as zero degree vertex.

Pendant Vertex: A vertex of degree one is called a pendant vertex or an end vertex. The vertex v_5 in Figure 10.1 is a pendant vertex.

Edges in Series: Two adjacent edges are said to be in series if their common vertex is of degree two. In Figure 10.1, e_3 and e_4 are in series. These two edges are incident on v_1.

Isomorphic Graphs

Two graphs G and G' are said to be isomorphic or equivalent to each other if there is one to the correspondence between their vertices and edges such that the incidence relationship is preserved. The isomorphic graphs are the same graphs drawn in different manner. The four graphs presented in Figure 10.2 are all isomorphic graphs, even though they look different because of different form. Of course, it is not always easy to know the isomorphism of two graphs, but isomorphism can be identified by the following characteristics:

1. They must have same number of vertices.
2. They must have same number of edges.
3. An equal number of vertices with a given degree.

Sub-graphs

A graph g said to be sub-graph of G, if all the vertices and edges of g are in G; and each edge of g has same end vertices in g as in G. A single vertex or a single edge together with its end vertices are also regarded as sub-graph of G. Within the graph if there are two sub-graphs and they do not have any edge in common, but isomorphic in nature they are called Edge-disjoint sub-graphs.

Walks

A walk is a finite alternating sequence of vertices and edges, beginning and ending with a vertex and each edge is incident with the vertices preceding and following it. No edge appears more than once, but a vertex may appear more than once. A walk is also referred as edge train or a chain. A walk is also a sub-graph of G. The vertices from which a walk begins and ends are called terminal vertex. Walks can be classified as closed walk and open walk. When a walk begins and ends at the same vertex, it is called closed walk, on the other hand if the two terminal vertices are distinct it is called a open walk.

Paths

An open walk in which no vertex appears more than once is called a path. A path does not intersect itself. Number of edges in a path is called the length of a path. A self-loop cannot be included in a path as it connects the same vertex twice by the same edge, but a self-loop can be included in a walk. An edge which is not a self-loop is a path of length one. The terminal vertices in a path are degree one and rest of the vertices in the path are degree two.

Circute

A closed walk in which no vertex appears more than once is called a circute. Hence a circute can be defined as a closed non-intersecting walk. The circutes are also called differently as cycles, circular path and polygon.

Matrix Representation of Graphs

Matrix is a convenient and useful way of representing a graph. Matrices can be easily analysed mathematically and matrix alzebra can be readily applicable to study the structural properties of graphs.

Incidence Matrix

A n × e matrix can be put at A= (a_{ij}) where the rows corresponding to n vertices and column for e edges. The matrix elements are shown in binary form, either in 1 or 0.

a_{ij} = 1, if ej is incident on vertex v_1 or
= 0 otherwise.

This type of matrix is known as incidence matrix. Since the incidence matrix contains only two elements, 0 or 1, it is also called binary matrix. Given a graph G, without any self-loop, it can be presented in the form of an incidence matrix. Some of the important characteristics of incidence matrix are:

1. Since every edge is incident on exactly two vertices, each column has exactly two 1's.
2. The number of 1's in each row equals the degree of the corresponding vertex.
3. A row with 0's represent an isolated vertex.
4. Parallel edges in a graph produce identical columns in its incidence matrix.

5. Permutation of any two rows and columns in an incidence matrix simply corresponds to relabelling the vertices and edges of the same G.

Accordingly a circute matrix and a cut-set matrix can be prepared. In a circute matrix, circutes are represented in rows while in cut-set matrix the cut-set are presented in rows but in both the matrices the columns are represented by the edges.

The incidence matrix of Figure 10.1 is presented below:

$$\begin{array}{cccccccc} & e_1 & e_2 & e_3 & e_4 & e_5 & e_6 & e_7 \\ v_1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ v_2 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ v_3 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ v_4 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ v_5 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}$$

Adjacent Matrix

When a connection matrix is representing a graph it is called adjacency matrix. The connections among the vertices are represented with no parallel edges. The adjacency matrix of graph G with n vertices and no parallel edges is a n × n symmetric matrix with binary notations of 0 and 1's. This matrix $X = (x_{ij})$ where

$X = 1$, if there is an edge between them or

$= 0$, if there is no edge between them.

The adjacency matrix below represents Figure 10.4.

$$\begin{array}{ccccccc} & v_1 & v_2 & v_3 & v_4 & v_5 & v_6 \\ v_1 & 0 & 1 & 0 & 0 & 1 & 1 \\ v_2 & 1 & 0 & 0 & 1 & 1 & 0 \\ v_3 & 0 & 0 & 0 & 1 & 0 & 0 \\ v_4 & 0 & 1 & 1 & 0 & 1 & 1 \\ v_5 & 1 & 1 & 0 & 1 & 0 & 0 \\ v_6 & 1 & 0 & 0 & 1 & 0 & 0 \end{array}$$

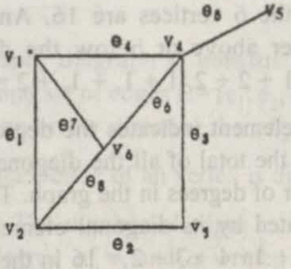


Fig. 10.4 Graph of an Adjacency Matrix

The main characteristics of adjacency matrix are:

1. The elements in the principal diagonals are O's if the graph has no self-loops.
2. If the graph has no self-loops or parallel edges the degree of vertex equals the number of 1's in corresponding row or column.
3. Permutation of rows and the corresponding columns imply the reordering of the vertices. If two rows are interchanged the corresponding columns must also have to be interchanged.

Power of Adjacency Matrix

Once an adjacency matrix X is prepared X^2, X^3 and X^n matrices can be computed. The 6 × 6 symmetric matrix representing graph shown in Figure 10.3 is presented below as adjacency matrix of the order of X^2, that is, matrix multiplying X matrix with the same matrix:

$$X^2 = \begin{matrix} 3 & 1 & 0 & 3 & 1 & 0 \\ 1 & 3 & 1 & 1 & 2 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 3 & 1 & 0 & 4 & 1 & 0 \\ 1 & 2 & 1 & 1 & 3 & 2 \\ 0 & 2 & 1 & 0 & 2 & 2 \end{matrix}$$

The X^2 matrix indicates the following information with regards to the Figure 10.4

1. The sum total of off diagonal elements indicate the total number of degrees of the vertices of G. In Figure 10.4 the total number

of degrees of the 6 vertices are 16. And the off diagonal elements (either above or below the diagonals) together 1 + 3 + 1, 1 + 1 + 2 + 2, 1 + 1, + 1, + 2 = 16.

2. Each diagonal element indicates the degree of the respective vertices, hence the total of all the diagonal elements indicate the total number of degrees in the graph. The v_4 has 4 degrees which is indicated by its diagonal element. The total of all elements 3 + 3 + 1+ 4 + 3 + 2 = 16 in the main diagonal.
3. The matrix also indicates number of different paths of length between *i*th and *j*th vertices. It can be noted that between v_2, v_5 the two connected paths of length two are: e_1 e_2 and e_4 e_5; which is indicated in X^2 matrix as $X_{25} = 2$.
4. The zeros indicate that there is no path of length between the two vertices.
5. The total of diagonal elements in X^2 (Figure 10.4) indicates the total number of positions in which matrix X has 1's. It can be noted that in matrix X there are total of 16 ones.

X^3 Matrix

By multiplying again X with X^2 matrix we get the X^3 matrix.

$$X^3 = \begin{matrix} 2 & 7 & 3 & 2 & 7 & 6 \\ 7 & 4 & 1 & 8 & 5 & 2 \\ 3 & 1 & 0 & 4 & 1 & 0 \\ 2 & 8 & 4 & 2 & 8 & 7 \\ 7 & 5 & 1 & 8 & 4 & 2 \\ 6 & 2 & 0 & 7 & 2 & 0 \end{matrix}$$

In the above matrix the off diagonal elements indicate the number of edge sequences of three edges between *i*th and *j*th vertices. The diagonal entry in X^3 equals twice the number of different circutes of length three in the graph passing through the corresponding vertex. For instance in v_2 and $X^3_{22} = 4$, this means there are two triangles in the graph which passes through v_2; that is, v_2, v_1, v_5 and v_2, v_4, v_5.

Directed Graphs

A directed graph, diagraph G consists of a set of vertices $v = (v_1, v_2, v_3 \ldots\ldots)$ and set of edges $E = (e_1, e_2, e_3 \ldots\ldots)$ where every edge is shown with an arrow directing the edge from v_i to v_j.

In a diagraph incidence of an vertex is divided into two:

1. *Initial Vertex:* The number of edges incident out of a vertex is called out-degree of v_i and is written as $d^+ (v_i)$. It is also known as out-valence or outward demi-degree. In Figure 10.4 for example, $d^+ (v_1) = 3$; $d^+ (v_5) = 4$.
2. *Terminal Vertex:* The number of edges incident into v_i is called the in-degree of v_i; and is written as $d (v_i)$. In the Figure 10.4 $d^-(v_4) = 3$ while $d^+(v_4) = 1$.

A vertex v in a diagraph is called pandent if it is of degree 1, that is if

$$d^+(v) + d^-(v) = 1$$

In a diagraph sum of all in-degrees are equal to the sum of all out-degrees of the vertices. A diagraph that has no self-loop or parallel edges is called a simple diagraph. Figure 10.5 shows a simple diagraph.

Incidence Matrix of a Diagraph

The incidence matrix of a simple diagraph where the rows correspond to vertices and columns correspond to edges, the elements are put in the matrix as follows:

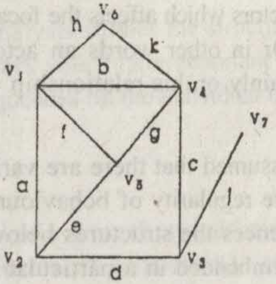


Fig. 10.5. A Simple Diagraph

The element a_{ij} = 1 if jth edge is incident out of v.

= –1 if incident into the vertex.

= 0 if there is no incident.

The incidence matrix of Figure 10.4 is shown below:

$$\begin{array}{c|cccccccccc}
 & a & b & c & d & e & f & g & h & k & l \\
v_1 & 1 & -1 & 0 & 0 & 0 & -1 & 0 & -1 & 0 & 0 \\
v_2 & 0 & 1 & 1 & 0 & 0 & 0 & -1 & 0 & 1 & 0 \\
v_3 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\
v_4 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & -1 \\
v_5 & 0 & 0 & 0 & 0 & -1 & 1 & 1 & 0 & 0 & 0 \\
v_6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\
v_7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
\hline
 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}$$

It can be found out that the column total of the incidence matrix are zeros since a particular edge has two connections of in-degree and out-degree on two vertices. The row total of 1's and –1's indicate total number of out-degrees and in-degrees and their sum total (ignoring the sign) indicate the degree of the respective vertices.

Sociogram

In the network analysis of social behaviour, or sociograms the emphasis is on social behaviour of the individuals in the society who are known as actors. An actor participates in a social system involving many other actors which affects the focal actor's perceptions, beliefs and actions. Or in other words an actor's socio-economic behaviour depends mainly on his relationship with other actors in the social circle.

Secondly, it is assumed that there are various levels of social structure, and there are regularity of behaviour in a given level of structure, which influences the structures below and above it. The individual actors are embedded in a particular social structure and they have no existence at the level of the individual actor.

Types of Analysis

Two basic approaches of analysing an actor's behaviour are: (a) according to their attributes and (b) relationships.

Attributes are the intrinsic characteristics of people, objects or events. Traditionally whenever a study is conducted it is mainly based on the attributes. A person's income, age, occupation, nations GNP or persons opinion are the attributes and they are measurable.

Persons, objects and events are also involved in relationships. Here the actions and qualities are considered together with two or more actors in a system. A relation is not an intrinsic characteristic in isolation, it is an emergent property of the linkage between the units of observation. The attributes of a person remain constant over a time, but the relationships are context specific. A student/teacher relationship does not exist after the schooling of a student. Among the students the relationship changes when they enter into different jobs. In these relationships the relational properties can be measured.

The attributes and relationships are conceptually distinct approaches to social research, but they should be seen neither polar nor mutually exclusive measurement options. Most of the social research to-day relies upon attribute measures through surveys and experiments. But relationship cannot be measured by simply aggregating the attributes of the individual members. Individual actor's attributes influence the social structure and the social structure in turn also affects the individuals. The ultimate advance of social scientific knowledge requires the combination of both the approaches for an accurate analysis of the system.

The network analysis studies the structure of relations among the actors in the network which have important behavioural perceptual and attitudinal consequences on the individual units and on the system as a whole.

Research Design

Research design of a network analysis involves four components: (1) Sampling (2) Forms of relations (3) relational content and (4) the level of data analysis.

Sampling

The earlier network analysts preferred only smaller samples because large scale samples posed complex calculations and there were no computers to help in the analysis as it is today. Even today most of the analysts take up small size samples. When a sampled actor is analysed in a network work population it ignores the non-sampled persons and thereby strong relations are isolated from the actual network. If the sample is 10 per cent that means, information with regards to 90 per cent of the network relations are lost. Those lost data are equally significant.

Potential linkages in a population size N, is N (N – 1)/2. Hence if 100 is taken as sample the possible linkages would be 4950. Therefore large scale samples pose difficulties in analysis.

Two of the important methods of network sampling theory are briefly discussed here.

1. Ove Frank as early as 1971, a Swedish mathematician explained the possibility of estimating the number of ties in a population size N, given a sample size n, which was further developed by Granovetter later in 1976. In this method a random sampling of actors is drawn and list is made of the names of respondents. Then the respondents are asked their acquaintances with the list of names and density of their network is established. The observed density of linkages between the respondents is taken as unbiased estimate of the density of the entire network.
2. A second method suggested by Beniger in 1976, is based on standard random surveys. In this method first the entire population is divided into classes according to the attributes of the population. In this sampling procedure information is collected from each respondent the attributes of persons to whom he goes for each type of relations studied. The combination of attributes are identified so that individuals falling into each category have similar relations with others.

For example, two important attributes of a person are race and age. Race can be white (w) and black (b) and age young (y) and old (o). The four combinations would be as follows:

	w	b
y	wy	by
o	wo	bo

z_j : Total number of persons that respondent j cites during interview.

f, wy : Number of citations to persons who are young and white.

The proportion of j's relations that involve persons white and young are:

$z_j, wy = f_j, wy/r_j$

Assume that respondent j is black and old. Now it is possible to sum relations to w, y people across all respondents who are b and o. That is

$$f_{(bo,wy)} = \sum_{j=1}^{n} (Z_{j,wy})(d_{j,bo})$$

where $d_{j,bo}$ is a dummy variable equal to zero unless respondent j is black and old. If no black and old respondent cites some one who is young and white then:

$f_{(bo,wy)}) = 0$

nbo indicates number of black and old respondents, Hence:

$z_{(bo,wy)} = f_{(bo,wy)}/nbo$

This is the mean percentage of black and old respondents citations that are directed towards who are white and young.

The total vector relations for respondent j can be expressed as:

$z_j = (z_{j,wy}; z_{j,by}; z_{j,wo}; z_{j,bo}z_{wyj}; z_{byj}z_{woj}; z_{boj})$

In this equation the first four elements in the right hand side indicates the relation from the respondents; while the last four elements indicate relations to the respondent. Individuals with the identical attributes are assumed to the recipients of identical relations.

The most crucial step in this sampling is that the identification of all relevant attributes that satisfy the population. When the relevant attributes are omitted, people will be defined as equivalent.

Forms of Relations

Relations can be divided into forms and content. Relational form refers to properties of the connections between the pairs of actors independent of specific contents. The two basic aspects of relational forms are:

1. The intensity or strength of the link between pairs of actors.
2. The level of joint involvement in the same activities.

Relational Content

Content refers to the substantive type of relation presented in the connections. The content is decided by the analyst. The nature of the contents may be borrowing, gossiping helping, supervising etc. It depend upon the nature of the studies undertaken.

Levels of Analysis

The relationships can be analysed at four levels of analysis as per the object of the study by the analysts.

1. *Ego-centric:* In the ego-centric network each individual actor is analysed with relationship with others. Each actor is described by the number, magnitude and other characteristics of its linkages with other factors. It strongly resembles attribute based research.

2. *Dyads:* The next higher level of analysis in a network is dyads or the actors analysed in terms of pairs. If the sample size is N, there are $(N^2 - N)/2$ distinct units of analysis at dyadic level. In this analysis it is enquired whether or not a direct link exists between the two actors. It explains variations in dyadic relations as a function of joint characteristics of the pair.

3. ***Triads:*** Triangular relationship among the actors is known as triads. In a N sample size, there are N/2 distinct triads formed by selecting each possible sub-set of three vertices and their linkages.

4. ***Complete Network:*** Beyond the triad level, the level of analysis is the entire network or the system as a whole. This has become one of the most popular level of analysis of networks in recent years.

Structure of Networks

A major use of network analysis is to uncover the social structure of a complete network system. A position or social role of an actor depends upon the pattern of relations that connect the empirical actors to each other.

The structural analysis uncovers the fundamental social positions, and complexity of network is typically simplified because it reduces a large number of N actors into a small number of M positions.

(i) ***Social Cohesion:*** Actors aggregated together into a position to the degree that they are directly connected to each other by cohesive bonds are called social cohesion. They are also known as 'cliques' or 'social circles'. Lesser percentage of direct ties means lesser social cohesion.

(ii) ***Structural Equivalence:*** Second method of identifying network positions is structural equivalence. Actors are aggregated into jointly occupied positions to the extent that they have a common set of linkages to the other actors in the total network system. The actors need not have any direct ties to each other.

Whether one has to identify the position of an actor either in the form of a clique or structural equivalence it depends on the nature of study but R.S. Burt prefers structural equivalence criterion for the identification of structural positions because of these advantages:

- It includes a broader range of type of sub-groups.
- It extends the scope of type of sub-groups in which homogeneity of attitude and behaviour can be expected.

- It can be subjected to statistical tests of goodness of fit.
- It provides a basis for sampling population network in large systems.

Indices for Networks

A sociogram of 11 tribal women is presented in Figure 10.6. The linkages are established on the basis of lending-borrowing relations of the women. The number in the vertex circle indicates the individual tribal women and the edges lending-borrowing relationship among the women.

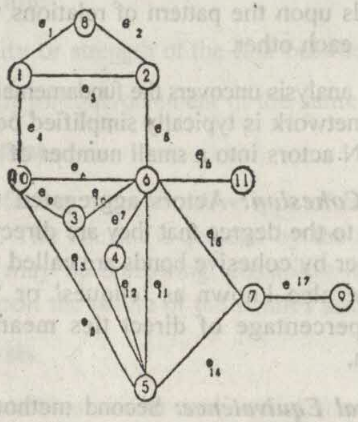


Fig. 10.6. Lending-Borrowing Activity Network

The network relations can be expressed in the form of a adjacency matrix. As it is an undirected graph the matrix would be symmetric and row or column total of 1's indicate the degree of the respective vertex or number of relationship of the actors, that is here the tribal women. This facilitates also the analysis of indirect relations. Indirect connections among a set of N actors in a network can be uncovered by raising an adjacency matrix to K and to successive powers.

The following is the adjacency matrix of lending-borrowing network of the tribal women shown in Figure 10.6.

$$M = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

By calculating M^2 and M^3 indirect reachability of an actor can be known. In M^2 one can know how two actors can reach each other through two links in the network, and in M^3 through three indirect links. In the above matrix M the total of 1's in the sixth row is 7, that means, the degree of the actor is 7, in other words the actor 6 can reach directly to 7 of the 11 network members.

The M^3 matrix is calculated of the above adjacent matrix and presented below. The elements in the matrix show the number of paths through which two members of the network can reach in three indirect linkages.

In M^3 matrix $M_{25} = 5$; which indicates that the network members two and five can reach each other in three step link in five ways. It can be verified that the five paths are:

(e_5, e_6, e_8) (e_3, e_4, e_8) (e_5, e_7, e_{12}) (e_5, e_{10}, e_{13})

(e_5, e_{15}, e_{14})

A reachability matrix can be computed R^T in the following manner:

$$R^T = M + M^2 + M^3 + \ldots\ldots\ldots M^T$$

$$M^3 = \begin{bmatrix} 2 & 6 & 3 & 3 & 3 & 3 & 3 & 4 & 0 & 7 & 2 \\ 6 & 2 & 1 & 1 & 5 & 10 & 1 & 4 & 1 & 3 & 0 \\ 3 & 3 & 6 & 4 & 11 & 13 & 4 & 2 & 2 & 8 & 2 \\ 3 & 1 & 4 & 2 & 9 & 11 & 2 & 1 & 2 & 4 & 1 \\ 3 & 5 & 11 & 9 & 10 & 13 & 10 & 2 & 1 & 12 & 4 \\ 3 & 10 & 13 & 11 & 13 & 10 & 12 & 2 & 1 & 15 & 7 \\ 3 & 1 & 4 & 2 & 10 & 12 & 2 & 1 & 3 & 4 & 1 \\ 4 & 4 & 2 & 1 & 2 & 2 & 1 & 2 & 0 & 2 & 1 \\ 0 & 1 & 2 & 2 & 1 & 1 & 3 & 0 & 0 & 2 & 1 \\ 7 & 3 & 8 & 4 & 12 & 15 & 4 & 2 & 2 & 6 & 2 \\ 2 & 0 & 2 & 1 & 4 & 7 & 1 & 1 & 1 & 2 & 0 \end{bmatrix}$$

The elements of R^T matrix are counts of the total number of linkages between the pairs of actors involving T or fewer steps. A zero element in the matrix means that the pair of actors not at all reachable.

A reachability matrix of M total up to three steps is compiled in the matrix R^3. This means $M + M^2 + M^3 = R^3$. In this matrix four elements are zeros. That means two pairs of actors cannot reach each other even within three links in the network. The two pairs not reachable within three steps are:

Actors 1-9, and 8-9.

For reachability they require four steps each as it can be verified by finding the linkages in Figure 10.6. The shortest path between 1 and 9 is e_4, e_6, e_{15}, e_{17}. (Of course there are other four edge paths in the network). Between actors 8 and 9 the path is e_2, e_5, e_{15}, e_{17}, and this is the single four edge path.

Path Distance

Reachability among the members of the network can be shown in a Path Distance Matrix. The matrix D^T is a path distance matrix. The elements in the matrix indicate the shortest chain linking actors i and j in the matrix.

The network diagram presented in Figure 10.6 is not a diagraph hence the 11 tribal women among themselves have maximum of 55 relations, and the matrix show these path distances:

	No. of pairs	Path Distance
	17	1
	23	2
	13	3
	2	4
Total	55	

$$D^T = \begin{bmatrix} 0 & 1 & 2 & 3 & 2 & 2 & 3 & 1 & 4 & 1 & 3 \\ 1 & 0 & 2 & 2 & 2 & 1 & 2 & 1 & 3 & 2 & 2 \\ 2 & 2 & 0 & 2 & 1 & 1 & 2 & 3 & 3 & 1 & 2 \\ 3 & 2 & 2 & 0 & 1 & 1 & 2 & 3 & 3 & 2 & 2 \\ 2 & 2 & 1 & 1 & 0 & 1 & 1 & 3 & 2 & 1 & 2 \\ 2 & 1 & 1 & 1 & 1 & 0 & 1 & 2 & 2 & 1 & 1 \\ 3 & 2 & 2 & 2 & 1 & 1 & 0 & 3 & 1 & 2 & 2 \\ 1 & 1 & 3 & 3 & 3 & 2 & 3 & 0 & 4 & 2 & 3 \\ 4 & 3 & 3 & 3 & 2 & 2 & 1 & 4 & 0 & 3 & 3 \\ 1 & 2 & 1 & 2 & 1 & 1 & 2 & 2 & 3 & 0 & 2 \\ 3 & 2 & 2 & 2 & 2 & 1 & 2 & 3 & 3 & 2 & 0 \end{bmatrix}$$

For comparison of networks with different densities and different number of members, Lincoln and Miller suggested a normalisation of path distances with this formula:

$$P_{ij} = 1 - \frac{d_{ij} - 1}{d_{max}}$$

where d_{max} is the longest path distance observed in the network between any reachable pair. The scores are calculated in the following manner:

1. When two actors are directly connected their path distance is 1, or the proximity value of the formula is 1.
2. When two actors are unconnected the path distance is 0, or $P_{jj} = 0$.
3. The other distances take the values ranging from 0 to 1, as per the formula given.

In the above path distance matrix, the maximum path distance is 4. Accordingly P_{ij} values of 2, 3, and 4 can be calculated:

$$P_{ij} = 2 = 1 - \frac{2-1}{4}$$

$$= 1 - 0.25 = 0.75$$

Accordingly 3 = 0.5; and 4 = 0.25. Hence the path distance matrix's elements other than 1's and 0's can be replaced by 0.75, 0.50, and 0.25 for 2, 3, 4 respectively to normalise the elements.

Relational intensity can be measured with the path distances and normalisation on an non-linear function to interpret decreasing relational intensity with increasing path distance.

Network Density

Density is a characteristic of a complete network. It is a proportion of all actual relations to the all possible relations. In calculating total relations the parallels and self loops are not taken into considerations. The density formula is:

$$\text{Density} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} Z_{ij}}{N^2 - N}$$

In the adjacency matrix representing the network relations of 11 tribal women shown earlier the actual relations are 34 while the possible relations are 110, hence the density of the network is

Density = 34/110
= 0.31

Network Cohesion

Cohesion is also an index of total network. The difference between density and cohesion is that in the density measurement asymmetric links are ignored while in the cohesion index asymmetric ties are taken into account. Network cohesion G, is a ratio of number of mutual choices in a binary matrix of directed links to the maximum possible number of such choices.

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} (z_i + z_j)}{(N^2 - N)/2}$$

In the equation G (z + z) is taken as 1 only if both the elements are 1s. Otherwise it takes the value of zero. Like density index G also ranges from 0 to 1.

Network Multiplicity

A link between the actors i and j is said to be multiplex if the same relationship is maintained in two or more networks among the actors, for the same set of N actors. For example, if actor i cites j in lending network, and same link is also maintained among the actors in another network, a network of say connections for collecting forest products together, they are said to be multiplex. Network multiplicity is of two types: (a) Network multiplicity and (b) Actor multiplicity. For the 11 tribal women a separate network of economic activity (collection of forest products) is prepared and presented in Figure 10.7. One can verify that the linkages are quite different from that of the lending-borrowing linkages in Figure 10.6. But we can find some similar linkages in both the networks. The network multiplicity can be calculated with the help of the following formula:

$$M = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} zij(m)}{N^2 - N}$$

z m) = 1 if both the networks or cross networks having same relations, otherwise it is zero. The multiplex network of two networks of lending-borrowing and economic activity is placed here in an adjacent matrix below:

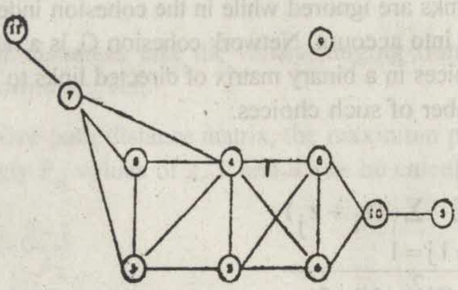


Fig. 10.7 Economic Activity Network

$$
\begin{bmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
$$

This multiplex matrix show that there are no common linkages with the actors 7, 9 and 11; hence the matrix is reduced to 8 × 8 adjacent matrix which is shown in Figure 10.8. According to the formula mentioned for the calculation of network multiplicity.

$$
\begin{aligned}
M &= \frac{18}{110} \\
&= 0.16
\end{aligned}
$$

Actor Multiplicity: Actor multiplicity can be measured as the proportion of an actor's linkages with all other N–1 actors.

$$M_i = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} z_{ij}(m)}{N-1}$$

In the multiplicity matrix $M_6 = 4$, hence the actor's multiplicity is 4/10 = 0.4. Similarly each actor's multiplicity can be calculated across two or more matrices.

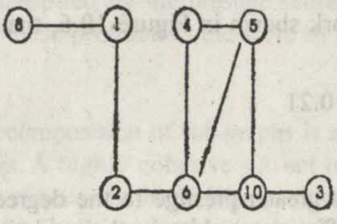


Fig. 10.8. Multiplex Network

Ego-Network Density

Each actor's ego-network density can be calculated to know the strength and status of the actor. It consists of that sub-set among N – 1 system with which the actor i has direct connection. This set is also called "first star" or primary star. If there are n changes in actors i's ego-network the density index is:

$$D = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} z_{ij}}{(N^2 - N)}$$

The Ego-density Index of network presented in Figure 10.8 can be calculated as:

D = 34/110 = 0.30

Centrality Index

Centrality index indicates the status of an actor in the network. More central an actor the greater his degree of involvement in all the network relations. The indices of centrality was first introduced by Bavelas in 1951. The simplest centrality index for actor i is a ratio of aggregate relations involving i over all relations in a network; that is, proportion of all network relations that involve i. The formula is:

$$C_i = \frac{\sum_{j=1}^{N} (z_{ij} + z_{ij})}{\sum_{i=1}^{N} \sum_{j=1}^{N} z_{ij}}$$

In the network shown in Figure 10.6, the centrality index of C_6, i.e. actor 6 is:

$$C_6 = 7/34 = 0.21$$

Prestige Index

An actor has greater prestige to the degree that other system members show difference to him in their relation. The prestigious actors receive more than initiative linkages, that is to be chosen rather than choosers. The simplest method to calculate this index is based on the in-degrees of an actor, or proportion of all these in the system that are directed towards an actor. But this method is possible for the diagraphs not for the general network.

Lin in 1976 first defined an actor's influence domain and provided an index of computation of prestige index.

$$P_i = \frac{I_i}{(C_i)(N-1)}$$

where I_i indicates the total number of members that directly or indirectly send relations to an individual. C_i is individual's centrality. It is the mean distance to all actors in the influence domain. If the centrality is zero, then a member's prestige is also zero.

Prestige index of actor 6 as per Figure 10.6 is calculated on the basis of P_i formula as under:

$$P_6 = \frac{.10}{(0.21)(10)}$$

$$= 4.8$$

There is another method of calculating prestige where both direct and indirect linkages are taken into account for the computation of prestige. In this method prestige of the rest of the actors' are also taken into account.

The formula for this is:

$$p_i = \sum_{j=1}^{N-1} p_j z_{ji}$$

This means each element in the column vector of the actor i in the matrix is multiplied by the prestige scores of the N-1 other actors and sum of these products is the score of i's prestige.

Clique Detection

Number and composition of sub-graphs is an important aspect of network analysis. A highly cohesive sub-set of actors within the network is called a clique. The implicit proposition motivating clique analysis is that the actors who maintains special cohesive bonds among themselves are more likely to perform similarly.

A clique can be defined as a set of completely linked sub-set which are directly connected to each other. A smallest clique consists of at least three actors each in mutual relation. Hence it is regarded as strongly connected sub-set. Another characteristic of a clique is that every pair of actors in a clique is adjacent.

In the network diagram shown in Figure 10.6 the following cliques can be identified, of which except the first one all others are overlapping cliques. If an actor is found in more than one clique it is called an overlapping clique.

The cliques are: (1, 2, 8) (5, 6, 10) (4, 5, 6) (5, 6, 7)
(3, 5, 6) (5, 3, 10) (3, 6, 10)

Structural Equivalence

Structural equivalence is another approach of partitioning the network members. This is widely used in comparing networks. Actors are aggregated into jointly occupied positions to the extent that they have a common set of linkages to other set of actors. This criterion requires a pair of actors to have exactly identical pattern of relations with N–2 other actors in the network, so that the two actors are placed in the same position in the network. Here a pair of actor's inter-relations are not taken into account, as in case of a clique, only their relation to other network actors are found out.

The structural equivalence of actors 5 and 6 in the network shown in the Figure 10.6 are given under:

Actor 5's linkages: 10-1-8-2; 3-10; 4; 7-9. The same linkages are also there for the actor 6 in the network, hence the pair of actors 5-6, are said to be having structural equivalence. Structural equivalence can also be shown in more than one network. It can be verified from Figure 10.6 and Figure 10.7 that there is structural equivalence of the pair of actors 5 and 6 in both the networks.

Actor 5's linkages in Figure 10.6:- 5-10-3; 5-4
Figure 10.7:- 5-10-3, 5-4

Same linkages are also there in both the networks for the actor 6. Both the methods of clique and structural equivalence can be applied to multiple network data, but more often cliques have been identified for the single networks while structural equivalence is applied to multiple networks.

Distance Measure

The structural equivalence is calculated on the basis of similar links with other systems actors; on the other hand distance is measured between the pairs of actors in terms of dissimilarity in their pattern of relationship with other system's actors. If the actors have identical relationship with other it is scored as zero. As the pairs have increasingly different pattern of linkages with others they are in more distance from others in the social space. Distance between pairs of actors with other network actors can be calculated as per the following formula:

$$d_{ij} = d_{ji} = \sqrt{\sum_{q-1}^{n} (z_{iq} - z_{jq})^2 + (z_{qi} - z_{qj})^2}$$

Distance between i and j; d_{ij}, equals the square root of the sum of squared differences across all third actors q. In the above equation $(z_{iq} - z_{jq})$ is the comparison between the two actors in the relations they initiate, i.e. the comparison of rows; while $(z_{qi} - z_{qj})$ is the discrepancy relations received from the third actors of the comparison of the columns.

The above formula for calculation of distance measures can be extended to more than one network of relations by summing values across K matrices involving same set of N actors in all the networks. For the distances involving more than one matrix the formula can be extended as:

$$d_{ij} = \sqrt{\sum_{K=1}^{K}\sum_{q=1}^{N}(z_{iq}-z_{jq})^2+(z_{qi}-z_{qj})^2}$$

Accordingly a matrix of social role distances can be calculated across the two or more networks and shown as Matrix D.

Let us calculate distance score between actors 1 and 5 across two networks of lending-borrowing and economic activity of the 11 tribal women shown earlier in this chapter. First place the row vectors of actor 1 and 5 parallel to each other taking the rows from the two adjacent matrices of the networks, and calculate the difference score.

	Lending-borrowing	Economic activity
1 :	0 1 0 0 0 0 0 1 0 1 0	0 1 0 1 0 0 1 1 0 0 0
5 :	0 0 1 1 0 1 1 0 0 1 0	0 1 0 1 0 1 0 0 0 1 0
Score	1 1 1 1 1 1	1 1 1 1 = 10

The difference score of column total will also be 10 as can be shown below:

	Lending-borrowing			Economic activity		
Actors	(1)	(5)		(1)	(5)	
	*	*		*	*	
	1	0	1	1	1	
	0	1	1	0	0	
	0	1	1	1	1	
	*	*		*	*	
	0	1	1	0	1	1
	0	1	1	1	0	1
	1	0	1	1	0	1
	0	0		0	0	
	1	1		0	1	1
	0	0		0	0	
		Total	6		Total	4

Total rows and columns differences in both the networks come to 10 + 10 = 20 and root over this comes to 4.47 as the distance score.

On the basis of the distance measures structurally equivalent positions of the actors can be identified; or the actors can be clustered. For the purpose of clustering a non-zero constant say k has to be set arbitrarily as

$$d_{ij} \leq k$$

Those actors having distance scores below or equal to k can be put in one cluster and above it as another cluster. By raising the k value gradually the successive clusters can be formed as in case of a hierarchical clustering method applied in case of cluster analysis.

11

Activity Analysis

One of the important application of networks is in project planning. Bar charts developed by Henry L. Gantt in the beginning of this century was the only method available previously for project planning. These work-versus-time graphs known commonly as bar charts become unusable for the large scale projects. After World War II when large projects were launched in the western world, serious difficulties were experienced in the management of the projects as the planning techniques used then felt inadequate to tackle the difficulties of such complexity and size.

Planners and managers were looking for a better scheduling and effective monitoring of projects. During the years 1956-58, two new techniques were developed in the USA known as Critical Path Method or CPM and Programme Evaluation and Review Technique or PERT. These two are almost similar in their approach and use the technique of networks for project planning.

Critical Path Method

The CPM was developed by a team of experts of the firms of E.I. Du Pont de Numours and Ramington Rand of USA in 1956. The Du Pont firm was facing difficulties in scheduling problems in their refinery renovation project. A team of mathematicians first worked out the general approach of the technique but Dr. J. Mauchly (UNIVAC Applications Research Centre) James Kelly (Ramington) and Morgan Walker (Du Pont) perfected the technique. The introduction of CPM for project planning saved Du Pont company to the tune of one million dollars. This method is now being extended to the field of cost and resource allocation. Planning Commission of India has been insisting the government to use this technique for the major projects.

Programme Evaluation and Review Technique

On the lines of CPM the US Navy independently was working to solve difficulties against a military background. In 1959 the US Navy faced problems in their "Polaris Missile Project" where they have to manage and control more than 3000 contractors and agencies to execute the project. The Navy set up an investigation team known as Programme Evaluation Research Task, hence the technique developed called as PERT. This technique basically shows the interdependencies between different events through a network diagram and identifies the key events which have to be reached to complete the project on schedule. The PERT now widely used by the engineers, economists and management scientists for managing monitoring and controlling the projects or organisations.

Network Planning

A network, as explained earlier, is a graphical plan which depicts the planned sequences of various operations for the achievement of project objectives. It is a logic model, as opposed to scale model, of the real system that seeks to represent. The networks, the CPM or PERT, in other words are the graphical portrayal of work plan in the form of networks and then numerical analysis to yield information of interest for the use of management. The planning here is mainly for the economy of time or to complete the project in minimum time possible and thereby saving of costs also. Delay in the completion of the project involves extra costs for the project in increase of cost of resources and wages and also shortage of supply in the economy if the project has to supply goods to the economy within a period of time

The network time planning process can be divided into three phases:

(1) Planning (2) Scheduling (3) Control.

First in the planning phase the main task is to identify the operations that are required to be performed for the completion of the project and establish proper sequential relationship between them. Three steps are mainly performed during this phase.

- *(i)* Breaking down of the entire project into suitable activities or various components of jobs for the execution of the project.
- *(ii)* Then the sequential inter-relationship is established among the activities.
- *(iii)* Finally a network diagram is drawn in the form of edges and vertices based on inter-relations. This is most important task of network analysis. The network here is a diagraph because the edges always indicate the direction towards which it flows.

Second, in scheduling phase, the detailed estimates are made with regard to each activity in the network for

- *(a)* the nature of job it has to be performed
- *(b)* time required for the completion of the activity, and
- *(c)* estimation of costs.

In calculating time of each activity, the network analyst finally have to identify the Critical Path, the actual time required for the completion of the project, and calculation of floats and slacks.

Third, the controlling and monitoring is as important as the other too. The project directors have to monitor the actual progress and the activities may be rescheduled as per the performance or due to arising of new situations or according to the availability of resources.

Drawing of Networks

In project planning networks only diagraphs are used and hence even if there are no specific arrow marks, it is assumed that the direction of lines are from left to right of the paper, and also from above to below. Some activities may be from lower to higher direction but they must be in clear arrow marks. Since these diagrams have arrow marks they are known also as arrow diagrams. The link in the form of arrow line or simple line is termed as activity or work-break, and the vertices which it connects are known as 'events'.

Activity: An activity is an element of work that has to be performed for the execution of the project. Each arrow or line is a unique one in the sense that an activity cannot be represented by more than one arrow. The beginning of the arrow or the tail indicates the start of the operation and the head or end of the arrow which touches the another event indicates its completion. Each activity starts and ends in an event. The name of the activity and the duration may be written on the activity line itself in full or using symbols. The time an activity takes for completion is known as activity time.

Events: The vertices in the network are called events. An event indicates a definite stage in completion of work, it represents a point of time during the execution of the project and it marks the start and finish of an activity. The events are shown generally in circles and are numbered. Squares and rectangles may also be used for the events. An event from which an activity starts is known as preceding event and at which it ends is known as succeeding event. The first event from which an activity starts is known as initial event. The initial event has zero time. This event can be named or numbered as 0 or 1. The last event of the network is called terminating event and its number indicates the total number of events in the entire network.

Designation of Activities

An activity can be designated by its events. It is designated in terms of its preceding and succeeding events. Preceding event is denoted as i and succeeding event as j. The succeeding event's number is greater than the number represented by the preceding event, or j is greater than i. For instance activity A starts from event 2 and ends in event 4; hence the description of the activity is A (ij) or A (24). A set of two event numbers indicate one activity in an unique way. An activity involves a definite amount of work, i.e. resources and money over time. Some activities which do not involve money or resources such as "Receiving goods" can also be treated as an activity. An event is said to be occur when all the activities leading to it have been completed. Except the first or the initial activity an activity is dependent on another activities and can start only after all the preceding activities have been completed.

Dummy Activities

Activities which are inserted in the network diagram to show the correct logic of sequential relationship between different activities or to ensure that the activities are uniquely defined as called dummy activity or simply dummies. The dummy activities do not consume either resources or time. It acts mainly as a "logic restraints" for representing the correct inter-relationship between the activities. Dummies, are known as zero time activities and shown as dashed lines with arrow heads. Dummies obey the same logic of precedence relationship as the normal activities.

A network diagram representing the following sequential relationship is shown in Figure 11.1.

A precedes B; B precedes E; A and C can go on parallel; D follows C; E can start only after B and D are complete; F follows D; G is the last activity and can start only after both E and F activities are completed. The dashed line connecting events 4, 5 is a dummy activity. It only maintains logic that the activity E can start only after the activities B and D are completed.

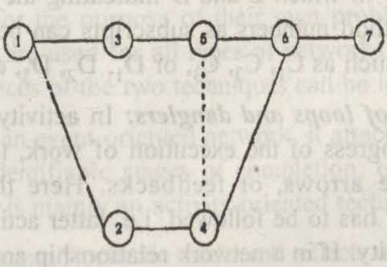


Fig. 11.1 Events and Activities

The Network Rules

Apart from the above methods of drawing an activity network, certain conventions are followed by network analysts in order to understand the networks by all concerned and interpret the network in clear and unambiguous terms. They are:

1. The time flows are set from left to right. That is the direction of the lines or arrows are from the left of the paper to the right of the paper. The arrows should be straight lines, and in bends acute angles are avoided or else the diagram shall be untidy, and difficult to read. The crossing of the arrows should be avoided as far as possible and if it is necessary minimum crossovers are used. Length and orientation of the lines are not important, it has relationship with the duration of the activity. It is desirable to ensure that the arrows are neither too long or too short to make a neater diagram. Without disturbing the relational logic it is also desirable to keep the dummy activities as minimum as possible.

2. Event numbers must be in ascending order from left to right, and there should not be any repetitions. It should be seen that the succeeding event numbers must always be greater than the preceding event numbers. Alpha numeric numbers can be used in big projects networks. In order to indicate the sub-sectors of the network concerning an important activity, alphabets or abbreviations as prefix can be used without breaking the continuity of the event numbers, as C1, C2, C4... Cn or D1, D2, D3 etc. in which C and D indicating the sub-sectors. In another method numbers as subscripts can be used for each sub-sector such as C_1, C_2, C_3, or D_1, D_2, D_3, etc.

3. *Avoidance of loops and danglers:* In activity analysis since it is the progress of the execution of work, there cannot be any reverse arrows, or feedbacks. Here the role of the dependency has to be followed, i.e. latter activity depend on earlier activity. If in a network relationship an earlier activity shown dependent on a later activity it is known as "loop" and a loop indicates an error of logic in the network. In a loop one has to go round and round, and there would not be any possibility of reaching the end of the network. The events 4, 5 and 6 makes a loop in Figure 11.2.

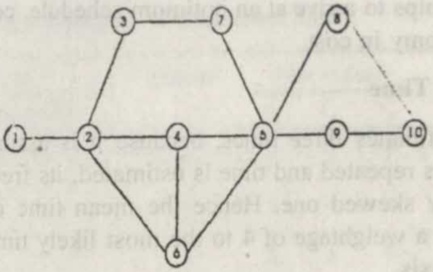


Fig. 11.2. Loops in a Network Graph

An activity that ends in a dead end, without completing the project or no activity follows from it, is called a dangling activity or danglers. If an event does not have any outgoing arrow a dead end will result. According to the network logic this is not permitted. Generally dangling activities creepes into a network when new activities are added in the revised network layout. If possible the dangling activity has to be connected with a dummy activity to keep the relational logic or else it should be removed and the network is redrawn to avoid the danglings. In the Figure 11.2 the event 8 is a dangler; hence a dummy activity has been added connecting the event 8 with the terminal event 10.

PERT and CPM

The PERT and the CPM even though developed by two different organisations for the purpose of their own problems, to-day both the techniques are used for all types of networks. But some of the minute differences of the two techniques can be identified.

1. PERT is an event oriented network, it attaches importance to certain identifiable stages of completion; on the other hand the CPM is mainly an activity-oriented technique.
2. PERT assumes stochastic model and reflects uncertainty in terms of completion of the activities. On the other hand CPM is a deterministic model. PERT uses three time estimates of activity duration while CPM takes only one time estimate.
3. In the earlier versions of PERT cost duration estimates were not used, but from the beginning CPM use cost duration

relationships to arrive at an optimum schedule, consistent with the economy in cost.

Estimation of Time

PERT estimates three times, because it is assumed that if a same activity is repeated and time is estimated, its frequency curve is a negatively skewed one. Hence the mean time of three time estimates with a weightage of 4 to the most likely time estimate is taken for analysis.

The three time durations are as follows:

(i) *Optimistic Time:* t_o: This is the shortest time interval required for the execution of an activity. If everything goes smoothly this time is reached.

(ii) *Pessimistic Time:* t_p: This is the longest time that is expected for the completion of the activity. This signifies a situation where everything goes wrong during the execution of the project.

(iii) *Most Likely Time:* t_m: This is the most likely time of the execution of the activity, that is, a high probability is attached to this time. This is taken as the mode of the frequency.

Since only one of the time estimates has to be adopted for the network a weighted mean is taken out of the three in the following formula:

$$\text{Expected time } t_e = \frac{(t_o + 4t_m + t_p)}{6}$$

Table 11.1 shows the estimation of expected time of eight activity, six event network presented in Figure 11.2. The subscript numbers of the activities indicate the expected time required for the completion of the respective activities.

Event Time

Every event has two times: earliest event time, TE and latest event time, TL. In a network since the event No.1 is the starting

Table 11.1: Calculation of Expected Time

Event Nos.		*Activity*	*Time in Days*			*Expected Time*
Predecessor	*Successor*		t_o	t_m	t_p	t_e
1	2	a	3	5	7	5
1	3	b	2	4	6	4
1	5	c	3	5	7	5
2	4	d	2	3	4	3
2	5	e	3	4	5	4
3	6	f	4	6	8	6
4	6	g	4	6	8	6
5	6	h	5	6	7	6

event, its TE is zero. To calculate TE of event 2 add 5 (activity a's expected time) to event 1's zero; i.e. 0 + 5 = 5. Accordingly TE of event 4 is 0 + 5 + 3 = 8; i.e. total time of activities a and d. Like this an event's TE is calculated by adding all the activity times earlier to it linked to the event in the network.

Every event has also a latest event time denoted as TL. It is calculated from the TE of the terminal event backward. It is assumed that TE = TL in case of the last event of the network. By proceeding backward from the terminal event 6 of the network under discussion, one can find out the TL of event 5. Activity h connects events 5 and 6; and event 6's TE is 15; hence by deducting activity h's t_e from TE of event 6 TL of event 5 is calculated.

TL of event 5 = 15 – 6 = 9 days

Calculation of TE and TL for the network shown in Figure 11.3 are presented in Table 11.2.

Slack Time

Slack time is the time by which the start of an activity taking off from an event can be allowed to be delayed without effecting the scheduled completion date. The slack time is the difference between the event times TE and TL:

Slack = TL – TE

Table 11.2: Calculation of Event and Slack Time

Event	*TE*	*TL*	*Slack time*
1	0	0	0
2	5	5	0
3	4	9	5
4	8	9	1
5	9	9	0
6	15	15	0

There are two important uses of slack time. The slack time enables the project executor to know the time at his hand in executing certain activities without affecting the total project time. In the table 11.2, two events viz. 3 and 4 have slacks. That means execution of the activity starting from the event 3 can be delayed for 5 days without affecting the project completion time. The slack time is also a good indicator of the critical path of the network. The slack time 0 in the table indicate that the events are in the critical path.

Finding the Critical Path

In a network there are several paths between the initial and the terminal events. In other words from the first event one can reach the last event of the network in several ways through the linkages of activities. Of the several paths the longest continuous chain of activities through the events, that is the longest path in terms of time, starting from the first to the last event, is called the Critical Path.

In the Figure 11.3 the network has four paths. The paths are:

1. Events 1, 2, 4, and 6.
2. Events 1, 2, 5 and 6.

3. Events 1, 5 and 6.
4. Events 1, 3 and 6.

Of these the longest path is the second one with a time of 15 days. Hence this path is known as critical path of the network. The total time of the critical path at the end of the event 6 is the minimum time estimate of the network to finish the entire project. Therefore isolation of critical path is most important in activity analysis.

Activities in the critical path are known as critical activities. Execution of the critical activities are very important because any delay in their execution would result in delay in the completion of the entire network. Activities other than in the critical path are known as non-critical activities. These activities can be delayed as per the slack time available without delaying the total project time. It should be noted that the events in the critical path have no slack time or their slack time is zero. The slack time zero implies that it does not have any spare time at all in executing the following activity and any delay in the completion of the activity would delay the entire project.

There are many ways adopted by the analysts to mention TE and TL along with the event numbers in the network. Some of the

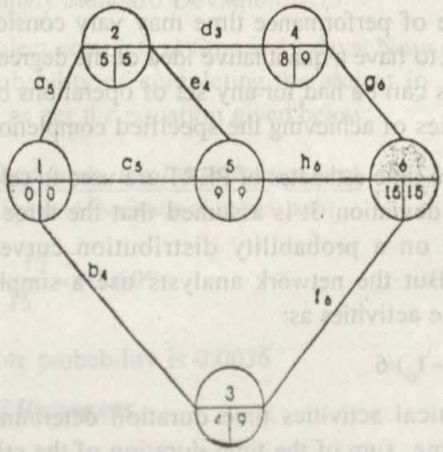


Fig. 11.3. The Critical Path

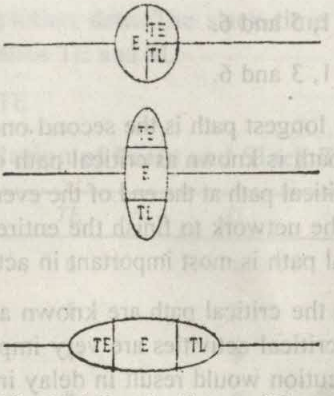


Fig. 11.4. Popular Forms of Events

popular forms adopted by the network analysts are presented in Figure 11.4. The critical path in the network is indicated by bold lines on the critical activities in order to distinguish from the non-critical paths. It can also be shown in colour lines. Similarly the events connecting critical activities or the critical events can be shown in double rings, colours or in shades.

Probabilities of Completion

When operation of a project depends upon large number of unknown factors, which are subject to change in an unpredictable way the value of performance time may vary considerably. Hence it is important to have a quantitative idea of the degree of uncertainty involved. This can be had for any set of operations by working out the probabilities of achieving the specified completion time.

The three time estimates of PERT are used to calculate variance and standard deviation. It is assumed that the three value of time estimates lie on a probability distribution curve having Beta distribution. But the network analysts use a simple formula for variance of the activities as:

$$V = (t_p - t_o)/6$$

Since the critical activities time duration determines the project completion time, sum of the time duration of the critical activities and their variances can easily be calculated. Let v_1, v_2, v_3, v_4....

v_n are the variances of n activities lying on the critical path. Hence the total of variances of the critical activities is the variance of the project:

$$V = \sum_{r=1}^{n} v_r$$

where v_r is the variance of the rth activity in the critical path. Based on the variance the standard deviation for the project completion time can be calculated as equal to the root over of the variance. Assuming the probability distribution curve for the project duration as nearly normal, probabilities of project completion time can be calculated.

Let us take the network provided in Figure 11.3 for calculation of probabilities for different project completion time.

Critical Events:	1	2	5	6
TE (days)	0	5	9	15
Critical Activity	a	e	h	
Variance	(4/6) (0.67)	(2/6) (0.33)	(2/6) (0.33)	

Sum of variance (0.67) + (0.33) + (0.33) = 1.33

Accordingly Standard Deviation: 1.15

The project completion time is 15 days. Now say we want to know the probability of completing the project in 12 days. It can be computed as per the equation given below:

$$F = \frac{\text{(desired time)} - \text{(project completion time)}}{\text{(Standard Deviation of critical path)}}$$

$$F = \frac{12 - 15}{1.15} = -2.609$$

Therefore probability is 0.0036

Allocation of Resources

Even though the planning of time is the main concern of the PERT analysis, but one cannot ignore the optimum use of resources for the project. The network should integrate economy in construction cost or organisational cost as per the availability of resources. Four types of resources are required for the operation of a project. They are:

1. ***Availability of Monetary Resources:*** Finance is the most important aspect of project execution. Timely availability of finance during the execution of the project enables the project to complete in time according to the network analysis. Any delay of the critical activities due to the non-availability of funds increases the duration of the entire project.
2. ***Human Resources:*** Manpower is another resource the project executive has to arrange for the timely completion of the project. Requirement of unskilled, semi-skilled and highly skilled labour has to be estimated and recruited as per specific need of the activities. Shortage of labour and non-availability of labour in time would increase the duration of time of the activities and hence the duration of project completion of time.
3. ***Material Resources:*** Apart from manpower a project requires material and machine. In a construction project week to week requirement of cement, sand, steel etc., has to be supplied regularly to execute the project as per schedule. Any shortage or irregular supply of materials would delay the activities beyond estimation and extend the period of completion of the project.
4. ***Machinery:*** Timely arrival of machines and equipment to be installed in the project is equally important for the completion of the project in scheduled time. The project director has to see that the required machinery are actually present in the project site for the execution of the activities.

Resource Scheduling

The main task in the activity analysis is linking of time schedules but these time schedules cannot be maintained without the scheduling of the resources. Even if there are no constraints with regard to the resources, they have to be smoothed for the project. Resource schedule prepared according to activities' demand fluctuates over time and it is difficult and uneconomic to arrange the resource as per the demand pattern. For instance, a particular skilled manpower requirement is scheduled as 10, 30 and 5 men over a period of three months. But one cannot employ them in this manner in a discontinued manner. Hence an average number of men are employed over a fixed period. This is called resource leveling. Of course this may result in over-employment of some of them and involve extra cost. Therefore it would be desirable to schedule the work in such a way that the resource requirement is spread more evenly over the entire project period.

Cost Analysis

There is a relationship between cost and duration of an activity. If a project has to be rescheduled for a shorter time more costs have to be incurred. In other words there is inverse relationship between cost and time; of course there is an upper and lower limit in the relationship. Whatever may be the high cost, a project cannot be completed below a certain minimum time. On the other hand whatever may be the time delay an activity's or a project's cost cannot be reduced below a certain minimum level. The project cost curve NN′ is non-linear and convex to the origin as shown in Figure 11.5. The p_c and p_n are the project crash cost and project normal cost

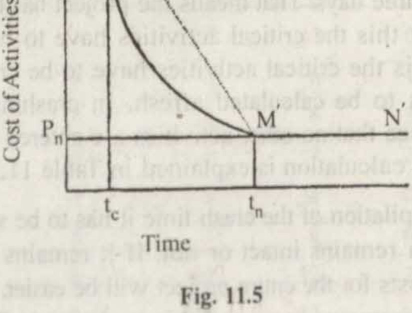


Fig. 11.5

respectively. The t_c and t_n are the crash or minimum and normal time respectively. The curve is non-linear because as one wants to reduce the duration of the project more than proportionate increase in the costs have to be incurred and vice-versa. But for an approximate calculation the relationship can be assumed as linear, which is shown in the figure 11.5 in the form of dashed line M M′.

Crashing of Projects

Advancing the date of completion by expediting some of the critical or other activities is known as crashing of projects. Crashing of a project may be necessary in two situations:

1. In course of the execution of the project some activities in the critical path might be already delayed. Hence to keep up the overall project completion time intact some of the critical activities require crashing.
2. The total completion time may require crashing because the firm may have to compete in the field or because of some over-riding considerations.

Cost of Crashing: The costs of activities in a network is generally worked out in normal costs. Reduction of time of completion requires fresh calculation of costs, both direct and indirect. This is because reduction of time involves more employment of labour, working of more costs and quicker supply of materials for the project.

In the network explained in Figure 11.3 the time required for the completion of the project is 15 days along with the events 1-2-5-6, which is the critical path. Now for instant the duration has to be crashed by three days. That means the project has to be completed in 12 days. For this the critical activities have to be completed in 12 days. For this the critical activities have to be crashed and cost of crashing has to be calculated afresh. In crashing activities the analyst should see that no other activities are entered into the critical path. The crash calculation is explained in Table 11.3.

After compilation of the crash time it has to be verified whether the critical path remains intact or not. If it remains intact then the calculation of costs for the entire project will be easier. In our example the critical path remains same and this is shown in Table 11.4.

The Table 11.4 shows that there is no change of the critical path as a result of the crash plan, the path 1-2-5-6 remained intact. But by reducing the crash time of the activities 'a' and 'e' the total completion time of the project time has been reduced to 12 days. The non-critical activities have also crash time, if they are not taken into account in the crashing plan there would be a different critical path; hence the non-critical activities crash costs have to be calculated to obtain the total crash costs. Table 11.5 shows the calculation of total crash costs for the project which includes crashing of both critical and non-critical activities.

Table 11.3: Calculation of Crash Costs

Activity	*Normal time (day)*	*Crash time (day)*	*Cost slope (Rs./day)*
a	5	3	60
b	4	2	40
c	5	3	50
d	3	2	30
e	4	3	40
f	6	4	20
g	5	5	—
h	6	6	—

Table 11.4: Recalculation of Critical Path

Tail Event	*Activity*	*Head Event*	*Finishing time*		*Slack*	*Remarks*
			TE	*TL*		
1	a	2	3	3	0	Critical
1	b	3	2	6	4	
1	c	5	3	6	3	
2	d	4	5	7	2	
2	e	5	6	6	0	Critical
3	f	6	2	8	6	
4	g	6	10	12	2	
5	h	6	12	12	0	Critical

Table 11.5: Calculation of Total Crash Costs

Activity	*Net Crash time days*	*Crash costs Rs. per day*	*Total cost of activity crash cost (Col.2 × Col.3)*
a	2	60	120
b	2	40	80
c	2	50	100
d	1	30	30
e	1	40	40
f	2	20	40
	Total crash cost of the project		410

It is evident from Table 11.5 that for reduction of three days in the duration of project completion time, a total of Rs. 410 has to be incurred extra. Crashing may involve changes in the critical path and accordingly the crash costs have to be compiled.

Control and Monitoring

Even though preparation of a network is perfect in all respects, the actual completion of a project depends upon effective control and updating the completion of the activities. Monitoring and updating must go simultaneously. As work or activity progresses there may be some difficulties in sticking to the activity time, hence it requires rescheduling the activities or crashing of some activities in order to complete the project in scheduled time.

The work progress can be shown visually in the network in different ways. The completed activities can be shown as double line arrows, in thicker lines or a colour line can be used over the original arrow lines.

12

Dimensional Analysis

Origin of dimensional analysis can be traced back from the early quarter of nineteenth century. In the year 1822, it was J.B.J. Fourier, a French mathematician first employed dimension as a concept of measurement. W. Stanlay Jevons was the first economist to apply this concept in economics in his book *"The Theory of Political Economy"* in 1879. A new branch of mathematics called mathematical analysis was developed by P.W. Bridgeman in his work *"Dimensional Analysis"*. Later Maurice Allsis, a French economist, presented a systematic theory of dimensions in economics in his book: *"Traite de Economie Pure"* published in 1943. Now the dimensional analysis is widely used in economics especially for the building of the models.

System of Measurement

In physical as well as in social sciences measurement forms primary aspect of tests, experiments and model building. In theorising and model-building the main aim is to establish relationship among the variables to be presented in the forms of equations. The value on both sides of an equation are measured in terms of quantities. And measurement of values is one of the important aims of a researcher.

Function of Measurement

There is a fundamental difference between counting and measuring. Counting is done through the pure numbers and expressed in non-convertible units. On the other hand measuring involves counting as well as units of measurement. In other words in counting pure number is always the case while the units of measurement change in a meaningful way.

There are two important functions of measurement. They are

1. It enables us to compare two different objects.
2. It enables us to find out the exact difference between the two.

When two individuals are compared, say Ram and Syam, expressing that Ram is taller than Syam; we make comparison between the two in measurement of height because Ram is few inches higher than Syam. Any comparison between small and big, high and low or short and long is conceived only through the measurement of the object.

Units of Measurement

Measurements are expressed in terms of an unit of measurement. The common units of measurements are grams, litres and metres. When one expresses measurement it is not simply an unit of measurement; it involves a pure number and an unit of measurement. When we say the weight of a good is 10 killograms, here 10 is a pure number and killogram is the unit of measurement. Hence any meaningful quantity is a number multiplied by an unit of measurement.

The term "dimension" is popularly used for geographical dimension or direction in space, usually in three directions of length, breadth and height. On the other hand 'dimensional analysis' is a mathematical technique. The Latin term *'dimensio'* means measuring. Hence a dimensional analysis is concerned with the theory of measuring different forms of values with the aid of mathematical techniques.

A single measurement unit of a quantity is known as a dimension. For instance weight is a dimension. Weight of an object can be expressed in terms of tonnes, killograms or grams. Accordingly length is another dimension. It can be expressed in terms of centimetres, metres or killometres. Time is also a dimension which is measured by minutes, hours or years. Since length is measured in one dimension (metres) and weight is measured in another dimension (killograms) these two dimensions cannot be added. For example, 10 metres cannot be added to 10 kg; but 10 metres can be added to 20 centimetres which are in the same dimension. Hence a dimension can be defined as a set of additive quantities.

Primary Dimensions

Dimensions are classified into two categories: (1) Primary dimensions and (2) Secondary dimensions. A quantity expressed in a single measurement is called a "primary dimension" or fundamental dimension. Most of the common measurements: length, mass, time etc., are the primary dimensions. The primary dimensions are simple and easy to understand. Measurements of sound, light, heat are also primary dimensions. In economics money, utility, resources in general are accepted as dimensions of fundamental type.

One can transform the number of measurement without changing its dimension. For example, money is a primary dimension whether in the units of rupees or U.S. dollars. An U.S. dollar is substitute for Indian rupees in the official exchange ratio of $1 = Rs.15. Indian currency of Rs.150 in one dimension can be expressed as $10. In this case the number 150 changed to 10 but in both the cases the dimension remained same. This can be explained in the following equation:

$$A = A$$

Here A and A' are the two different forms of measurement of money. If A = US dollars, A' = Indian rupees, hence both are in same dimension because they are the measurement of money in their respective countries. In both the sides of the equation lamda is the pure number which changes when US dollars are converted in Indian rupees but does not change the dimension.

Secondary Dimensions

Measurements involving two or more fundamental dimensions are called secondary dimensions of derived dimensions. Speed of an engine is expressed or measured in "killometres per hour". This is a derived dimension. Here the two fundamental dimensions of length (in Kms.) and time (per hour) are involved.

In economics price and demand are the secondary dimensions.

Price: Price is a derived dimension since it involves two dimensions, i.e. money and commodity. When we express the price of rice, Rs. 5 per kg., rupee is one primary dimension and kg. is another dimension. If simply it is said Rs. 5 it is a meaningless term.

Demand: Demand also is a secondary dimension. It involves two dimensions: quantity of goods which is demanded in kgs and a specific period of time over which it is demanded. Weight and time are the two fundamental dimensions. Without involving time, i.e. without expressing the period over which the quantity is demanded in the market, the expression of demand in quantity is meaningless.

A derived dimension may also involve more than two fundamental dimensions. For example, for the comparison of efficiency by railway transport, one can calculate the cost of carrying one kg. of goods per km. In this secondary dimension three fundamental dimensions are involved: Monetary unit, the weight of the goods, and length of track in kms.

In economics mostly derived dimensions are used. But unlike the derived dimensions of the physical sciences in economics they are not the standard units. Price is a derived dimension but it cannot be a standard measuring tool since the price itself changes over time. But a "r.p.m." (revolutions per minute) is a standard measurement used in physics. Time is an important fundamental dimension used both in physical as well as economic sciences widely.

Use of Algebra

Theorising and model-building involves equations and hence the use of algebra is inevitable in economics. Algebra can be divided into two forms: (1) classical algebra and (2) physical algebra.

Classical algebra or the ordinary algebra deals with mathematical or logical relationships; but physical algebra expresses relationships between things and not simply between two unknown entities. The linear equation.

$$Y = 10 + 8x$$

explains a relationship between the variables Y and X. It is an example of classical algebra; but it does not make any physical sense. Y and X must be assigned certain dimensions if at all the equation has to be used to explain any real relationship of objects. Classical algebra is the basis of physical algebra, but it is the latter which is of any practical importance in physical or social sciences.

Classical algebra can be distinguished from the physical algebra on the following points.

1. Classical algebra deals with the mathematical relationships between the pure numbers; while in physical algebra the relationships are explained between the things in terms of the equation. In the classical algebra either the things are not involved or they are manipulated on imaginary things.
2. In classical algebra no physical units of measurements are assigned to its variables; while in physical algebra all the quantities and numbers are assigned physical counterparts on both sides of an equation. In other words all the values are assigned dimensions on both sides of the equation.
3. Equations are solved, in classical algebra with reference to the rules of ordinary arithmetics without taking into account the underlying values of the variables, unlike the physical algebra.

Secondary Dimensions

Secondary dimension (S) is a product of primary dimension (P). We know about the transformation of primary dimensions in which the dimensions remained constant even though the constant pure numbers changed. In the transformation of primary to secondary dimension both the pure number and dimension changes. The dimensional equation for the secondary dimension can be written as

$$[S] = [P_1, P_2 \dots P_n]$$

The right hand side of the equation is called dimensional function. It should be noted that:

1. Dimensional function is always a product or power of the primary dimension.

2. The transformation equation and dimensional equation are the same in form but not in substance. The former expresses a functional relationship between numbers but the latter defines a functional relationships sets of quantity.

Fundamental Propositions

Dimensions are expressed in capital letters within square brackets. Dimensional analysis is mainly based on the following propositions:

Proposition No. 1

If a ε (read is an element of) (M) and b ε (M) then a ± b ε (M). This proposition explains that the variables associated with the same dimensions can be added or subtracted. Example:

[M] + [M] = [M]

Rs. 5 + Rs.10 = Rs. 15

Rs. 10 – Rs. 5 = Rs. 5

In other words two different dimensions, primary or secondary, are not possible for addition or subtraction, dimensionally. For instance quantity of a good say 10 kg. expressed as dimension [R] cannot be added or subtracted to [M], the dimension of money.

[M] + [R] = [?]

Rs.10 + Kgs.10 = ?

This is dimensionally inconsistent.

Proposition No. 2

If a ε [R] and b ε [M] then ab ε [RM] and a/b ε [RM^{-1}]. This proposition states that when two fundamental dimensions are associated in the form of multiplication or division the result become a derived dimension. In this proposition [RM] and [RM^{-1}] are the secondary dimensions.

Proposition No. 3

In a ε [L] and b ε [L] then ab ε [L^2]

Here a and b belongs to the same dimension. Hence the derived dimension is $[L^2]$; L is the dimension for length. So $[L^2]$ is the measurement for area in square metres. Accordingly volume can be measured in qubic metres as $[L^3]$ a derived dimension.

Proposition No. 4

If $a \varepsilon [M]$ and $b \varepsilon [M]$ then $a/b \varepsilon [1]$.

This shows that if one dimension is divided by the same dimension the result is [1]: that means the product is dimensionless. Sometimes dimensionless quantities are also used, that is, pure numbers are used for the purpose of comparison with a meaningful base. In economics the pure numbers such as elasticity of demand, index numbers and coefficient of correlation etc., are used as pure numbers which are dimensionless.

For example, Rs. 10/Rs. 10 = 1. Here the result is simply a pure number and conveys nothing in particular.

Proposition No. 5

If $Y \varepsilon [R]$ and $Y = f(x_1, x_2, x_3 \ldots\ldots x_n)$

then $f(x_1, x_2, x_3 \ldots\ldots x_n) \varepsilon [R]$

This proposition states that if two sides of an equation are equal then besides having same numerical value they must be same in dimension. This means the equation must be dimensionally homogeneous.

Type of Equations

Equations play crucial role in building theoretical models. In specifying relationships between economic variables one not only have to see mathematical rules but also to analyse the phenomenon must represent the point of view of economic specifications. Accordingly equations can be classified in economics into three categories.

1. Behavioural Equations
2. Institutional Equations.
3. Technical Equations.

Behavioural Equations

Behaviour of consumers in relations to other factors can be expressed verbally. A hypothesis about a behaviour can be expressed verbally but more often an equation is used to express the relationship because of economy, precision and hopefully for easier comparison. Most of the equations used in economics are behavioural equations.

Observe the following two equations:

1. $Q = f(P)$
2. $C = a + b(Y - T)$

These two equations are behavioural equations. The first one embodies the hypothesis that as the price of a commodity falls the consumers will plan to consume more. This is a statement about the behaviour of the consumers. The second equation explains a macro-economic behaviour. It explains the consumption function. Here the hypothesis is that the consumption is a linear function of disposable income. $(Y - T)$ is income minus taxes, indicate disposable income, C being consumption.

The second equation actually embodies two hypotheses. First, the consumption is related to disposable income; higher the disposable income higher would be absolute consumption. Secondly the relationship between the two is proportional or linear.

A mathematical form can be used to express behavioural relationships but the precise form that it takes and the constraints that are placed on the parameter values are determined by the economic considerations.

The above consumption function model indicates several economic considerations:

1. Consumption is an increasing function of disposable income. This means b is greater than 0.
2. The slope of the consumption function which measures the marginal propensity to consume lies between zero and unity: $0 < b < 1$.

3. The intercept 'a' is greater than zero: $a > 0$. This economic condition says that the expenditure can exceed the current disposable income. Here it is assumed a short period behaviour of the consumer, because the consumers can finance their expenditure out of their past savings.

Many equations theorising economic concepts does not imply behaviourism. Even though hypotheses can be put in an equation form, in reality there may not be any counterpart in the real world.

Institutional Equations

Certain equations are based on institutional factors, Government's policy of taxation, rate of foreign exchange and Central Bank policy of credit and money rates cannot be ignored in formulating a model which influences the behaviour of the consumers as well as the production system of the country.

General formula for the compound interest and funds available at the end of period is:

$$S = P\,(1 + r/100)^{t}$$

Here the future expenditure at the end of period (t), with an initial deposit of (P) gets (S). But this is dependent on the institutional r, the rate of interest given by the banks. Hence the behaviour of the consumers with regard to the deposits and future spending depends on the institutional factors.

S: Expenditure of future money. Money over time (MT^{-1})

P: Present savings stock of money: (M)

r: Rate of monetary interest charged over time:

$$\left[\frac{M/M}{T}\right] = [T^{-1}]$$

Replacing the original symbols with dimensions we get:

$$[MT^{-1}] = [MT^{-1}]$$

This equation is dimensionally homogeneous.

Technical Equations

Technical equations neither explains a behaviour pattern nor dependent on the institutional factors. Production function is a fine example of this form of equations.

$X = f(L, K)$

Which states that the production is the function of both labour (L) and capital (K). The Cobb-Douglas production function is most popular in economics which specifies production system in the field of manufacturing industries. The equation in the general form is:

$Q = A L^a K^b U$

Where Q is output, A is a constant for scale efficiency parameter, a and b are the positive constants, and the U is a random disturbance term. The numerical value of the exponents indicates the nature of the production function. The sum of the exponents show the degree or 'return to scale' and in the above function it is assumed as unity i.e. $a + b = 1$ or constant return to scale.

The above technical equation indicates the following:

(a) The input and the outputs are in the physical terms.

(b) The relationship is governed by the technology.

(c) It does not refer to human behaviour.

(d) Marginal and average products can be calculated.

Objectives of Dimensional Analysis

A model prepared in the form of a mathematical equation in economics have to be utilised in practice. For the purpose of testing whether the model is of any practical importance or not it should be tested through a dimensional analysis. A model is practicable only when the two sides of an equation are dimensionally homogeneous. This is one of the main objectives of dimensional analysis.

Let us test the following equation:

$Y = aX$

As a mathematical equation it is alright. But once we use it for the purpose of monetary analysis assigning specific values to the variables a dimensional analysis has to be made to test its validity.

If Y is the stock of money and X the flow of money and 'a' being a constant, this equation is wrong, since the flow of money cannot be equal to the stock of money.

Dimensionally $Y \varepsilon [M]$ and $X \varepsilon [MT^{-1}]$

Hence $[M] = [MT^{-1}]$

Since the dimensions of the both sides of the equation are different the equation is not dimensionally homogeneous and hence the equation is wrong.

Second, dimensional analysis offers a convenient device for isolating the source of error. The inconsistence of the equation mentioned above can be removed by introducing time elements in the variables of the left hand side of the equation; or by removing the time element from the right hand side of the equation. Hence dimensional analysis helps in discovering its economic meaning if this is not already explained by the equation.

Third, dimensional analysis presents a method of specifying if only partially, the forms of functional relationship between a number of variables. In other words dimensional analysis is not capable of completely defining the unknown functional relationships.

In the following equation:

$Y = aX^2 + bX + c$

When the values of the constants are not known the functional relationship is partially specified.

Fourth, dimension of certain variables tell us how the numerical value of those variables change when the units of measurement are subject to change.

Dimensional Entities

A pure number is a dimensional entity. When the result of an equation is a pure number the equation is regarded as dimensionless and is unsuitable for practical use. But in economics dimensionless entities are also in use and they are necessary for evaluation of certain problems.

Elasticity of demand is a pure number i.e. it is dimensionless. It is used to measure the degree of elasticity of demand.

Price elasticity is defined as:

$$e_p = \frac{dQ}{dP} \times \frac{P}{Q}$$

We know that demand ε $[RT^{-1}]$ so Q and dQ ε $[RT^{-1}]$

Price ε $[MR^{-1}]$ so P and dP $[MR^{-1}]$

$$\text{Hence: } \frac{dQ}{dP} = \frac{[RT^{-1}]}{[MR^{-1}]} = [R^2M^{-1}T^{-1}]$$

$$\frac{P}{Q} = \frac{[MR^{-1}]}{[RT^{-1}]} = [MTR^{-12}]$$

$$e_p = [R^2M^{-1}T^{-1}][MTR^{-12}] = [1]$$

Price index is another field where a pure number is used to compare changes of value over time. Fishers ideal index number formula is:

$$\text{Price Index: } \frac{P_nQ_n}{P_oQ_o}$$

Where the subscripts n and o indicate the current year and base year respectively for the quantities and prices.

Dimensionally:

$$\text{Price Index: } \frac{[MR^{-1}]}{[MR^{-1}]} - \frac{[RT^{-1}]}{[RT^{-1}]} = \frac{[MT^{-1}]}{[MT^{-1}]} = [1]$$

Dimensional Constants

In an equation when there are certain pure constants which are dimensionless, how it is accepted in a dimensional analysis? If a pure constant is withdrawn the meaning of equation will become wrong. So how to analyse the pure number in a dimensional analysis?

Let us take the equation of linear demand curve, in which P is price and Q is quantity demanded by the consumers:

$$P = a - bQ$$

In order to have a dimensional analysis of this equation the dimension of the coefficients 'a' and 'b' have to be determined.

The coefficients are pure numbers hence they are dimensionless. But the dimensions can be found out for these constants. Let us take 'b' first. In the equation it indicates the slope of the demand curve or it indicates the rate of change between P and Q, the two variables in the equation. Hence:

$$b = dP / dQ = \frac{[MR^{-1}]}{[RT^{-1}]} = [MTR^{-1}]$$

The coefficient 'a' even though it is a pure number a dimension can be found out for it. From the above equation the value of 'a' is:

$$a = P + bQ$$

Hence dimensionally:

$$\begin{aligned} a &= [MR^{-1}] + [MTR^{-2}]\,[RT^{-1}] \\ &= [MR^{-1}] + [MR^{-1}] \\ &= [MR^{-1}] \end{aligned}$$

when dimensions are found for the constants in an equation it is known as dimensional constant. Hence a dimensional analysis of the above equation shows that it is dimensionally homogeneous.

$$P = a + bQ.$$

Dimensionally:

$$\begin{aligned} [MR^{-1}] &= [MR^{-1}] + [MTR^{-2}]\,[RT^{-1}] \\ &= [MR^{-1}] + [MR^{-1}] \\ &= [MR^{-1}] \end{aligned}$$

Now observe the following equation:

$$P = a + Q$$

Dimensionally this equation is:

$$[MR^{-1}] = [MR^{-1}] + [RT^{-1}]$$

According to proposition No.1 the two dimensions in the right hand side of the equation cannot be added and hence it is not dimensionally homogeneous. This equation indicates that something is missing. Here the rate of change of Q is missing. If the slope 'b' defined as $b = 1$, which is implicit in the equation, it can be made dimensionally homogeneous.

Production function: The constant A in the Cobb-Dougles production function is also a dimensional constant:

$$Q = A\,L^aK^b \qquad (a + b = 1)$$

L : Employed manpower per unit of time R_1/T

K : Stock of capital R_k

A : Technical knowledge or the state of the art constant. Of course it is assumed constant but it can also change over time.

$$Q \,\varepsilon\, [R_eT^{-1}]$$

$$L \,\varepsilon\, [R_1T^{-1}]$$

$$K \,\varepsilon\, [R_k]$$

Hence $A = Q/(L^a.K^b)$ Where $b = 1 - a$

Dimensionally:

$$A = \frac{[R_eT^{-1}]}{[R^a{}_1T^{-1}][R_k{}^b]}$$

$$= \frac{[R_eT^{-1}]}{[R^a{}_1R^b{}_kT^{-a}]}$$

$$= [R_eR_1{}^{-a}R_k{}^{-b}T^{-b}]$$

This shows that 'A' here is not a pure number and dimensionless. These types of constants are known as dimensional constants. But caution may be taken that a constant in every equation cannot be treated as dimensional constant. Constant appearing in fundamental equations can be accepted as dimensional constant after finding its dimension.

Dimensional Analysis as an aid to memory

Certain equations are vaguely presented for operations keeping some concepts implicit in it, but for the purpose of practical operations a dimensional analysis indicates it in proper form.

The terms of trade between two countries in physical terms is explained in a simple equation where the subscripts A and B indicate the first and second country respectively:

Terms of Trade: Q_B / Q_A

Here Q_B is A's rate of import and Q_A is A's rate of export. Dimensionally the terms of trade is:

$$\frac{[R_B T^{-1}]}{[R_A T^{-1}]} = [R_B R_A^{-1}] \qquad (1)$$

Price level of the goods are indicated as P_A for country A's goods and P_B for B's goods in transaction. Dimensionally:

$P_A = [M_A R_A^{-1}]$

$P_B = [M_B R_B^{-1}]$ Expressed in their respective monetary units.

$$\text{Hence: } \frac{P_A}{P_B} = \frac{[M_A R_A^{-1}]}{[M_B R_B^{-1}]} = [R_B R_A^{-1} . M_A M_B^{-1}] \qquad (2)$$

The right hand side of the dimensions are not equal to the dimensions of the equation No.1 above, hence it is not dimensionally homogeneous.

What is missing in the equation? The missing factor is the rate of exchange of money between the two countries. Hence the rate of exchange of B's unit of money by A's unit of money has to be introduced in the equation to set it right.

The rate of exchange of money

$n_{AB}\varepsilon = [M_A M_B^{-1}]$

So the actual terms of trade should be:

$$\frac{P_A}{P_B \cdot n_{AB}}$$

Dimensionally the terms of Trade

$$= \frac{[R_B R_A{}^{-1}][M_A M_B{}^{-1}]}{[M_A M_B{}^{-1}]}$$

$$= R_B R_A{}^{-1}$$

It is now become homogeneous. The dimensional homogenity of an equation is not the proof that the equation is correct, but a theoretical equation which is not dimensionally homogeneous must be false.

Buckingham's Theorem

E. Buckingham in his book *"On Physically similar systems, Illustrations of the use of the Dimensional Equations"* published in 1914 made an good attempt to formulate principles for dimensional analysis. His theorem is still regarded as the main theorem of dimensional analysis.

The theorem: "Every dimensionally homogeneous equation can be written as a functional relationship between a complete set of mutually independent dimensionless products."

The main conditions of the theorem are:

(a) The equation should not state more than one relationship between the variables at the same time.

$Y^2 + Y + aX + bZ = 0$

This equation is not dimensionally homogeneous. Here the functional relationship among the variables is more than one. To make it homogeneous the equations can be split into two functions as $Y + aX = 0$ and $Y^2 + bZ = 0$

(b) Dimensional products belongs to one and same set of such products are said to be "mutually independent" if

no dimensionless product of this set can be expressed as product of powers of other dimensionless products of the set.

Say X/Y and X/Z are two dimensionless products. Hence it is also concluded that Y/Z is also a dimensionless product. But X/Y and X/Z are not mutually independent because in both the products there is one common element X even though they are dimensionless.

(c) A set of mutually independent dimensionless product is said to be "complete" if every other dimensionless product of variables of the equation can be expressed as a product of two or more dimensionless products belongs to the set.

So the question is, how to find out the maximum number of mutually independent dimensionless products that can be formed from the variables from a dimensionally homogeneous equation? A rule of the thumb formula is that the maximum number of independent dimensionless products is equal to the number of variables (n) less the number of primary dimensions (r) applying to these variables. Both dependent as well as independent variables are included in 'n'. The dimensional constants are also accepted as variables and included in 'n'.

Let us take Fisher's equation for analysis:

$$MV = PT$$

The dimensions for these variables are:

$$M\,\varepsilon[M] : V\,\varepsilon[T^{-1}];\; P\,\varepsilon[MR^{-1}];\; T[RT^{-1}]$$

In this equation there are four variables: M, V, P, and T. Hence $n = 4$. There are three dimensions: R, M, and T. So $r = 3$, hence $n - r = 1$; that means there must be one dimensionless product in the equation.

Next find out the dimensionless product by introducing one dimension after another equating it to a dimensionless product.

$M = b$ (an abstract number) hence $M\,\varepsilon\,(1)$.

To cancel M, dimension of P is introduced:

$$\frac{M}{P} = [1] \text{ or } \frac{[M]}{[MR^{-1}]} = [1]$$

Next we want to wipe out R hence T is introduced in the denominator.

$$\frac{[M]}{[MR^{-1}][RT^{-1}]} = [1]$$

Now T can be wiped out in the denominator by introducing V in the numerator:

$$\frac{[MT^{-1}]}{[MR^{-1}][RT^{-1}]} = [1]$$

$$\therefore \frac{MV}{PT} = b$$

Hence MV = b. P T

This shows that Fisher's equation MV = PT even though dimensionally consistent a dimensional constant is missing. It is impossible to determine the value of b. Economic theory has to specify it. If b is specified as b = 1, or in the dimensional analysis n – r + 1 formula is taken the equation becomes dimensionally consistent. The constant b here is important because P the price is not simply a price index. Hence the value of b has to be set on the basis of economic theory or empirical research. Algebraically better methods does not always yield better results from the point of view of economic theory.

Dimensional Analysis of Some Economic Theories

Marxian Theory of Production

Karl Marx in his *'Das Capital'* explained the theory of surplus value in terms of reproduction schemes as in the following form:

$$C + V + S = W$$

C and V are the constants and the variable capitals respectively and S is the surplus value created by the two former values. In the usual sense of these terms, the equation is not dimensionally homogeneous since C and V cannot be added as they are in two different dimensions. But to overcome this difficulty Marx explained them in the form of labour units: the socially necessary labour units. Hence C (the constant capital as past labour) and V (the living or working labour) and S the surplus value also in terms of labour units; are all in one single dimension, in labour units, hence there is no problem of dimensional inconsistency.

What Marx has done was that he explained all other commodities through a 'numeraire commodity' i.e. all goods and services defined through another commodity. Hence his equations are dimensionally homogeneous. For the purpose of dimensional analysis in economics when there is inconsistency, all other commodities can be expressed in terms of a single commodity and make the equation dimensionally consistent. In economics most commonly used numeraire is money.

Computation of a numeraire commodity can be explained with the help of the budget line of a consumer. In Figure 12.1; A-B line is the budget line, and x_1 (Rice) and x_2 (cloth) are measured in y and x axes respectively. For a given income of Y a consumer can purchase x_1 and x_2 as per the price of the commodities P_1 and P_2 respectively. Hence the budget equation is:

$$Y = P_1X_1 + P_2X_2$$

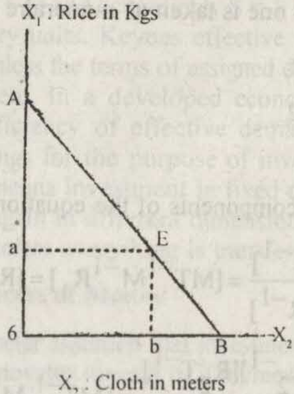


Fig. 12.1. Transformation to Numeraire Commodity

In the above equation the commodities are converted into monetary units, hence there is no difficulty of dimensional inconsistency, in physical sense x_1, x_2 cannot be added because rice in kg. and clothes in metres are in two different dimensions.

To avoid the inconsistency rice and cloth can be expressed in terms of a single commodity, i.e. rice taking as a numeraire commodity. In the given budget line at the point E the consumer purchases 'a' of x_1 and 'b' of x_2 respectively. The budget line is drawn on the basis of the prices of the two commodities. If the entire amount of money income Y is spent on purchase of x_2 only the consumer can buy OB, i.e. Y/p_2 = OB. Similarly the consumer can buy maximum of OA if he spends the entire income on x_1. Since the price of both the commodities are known, equivalent amount of x_1 can be found for the b_1 of x_2. This means rice becomes a numeraire commodity and adding of rice and cloth in terms of rice is dimensionally consistent.

Dimensionally demand for rice and cloth are:

$$x_1 \ \varepsilon [R_1 T^{-1}] \text{ and } x_2 \ \varepsilon [R_2 T^{-1}]$$

and price of the respective commodities

$$p_1 \ \varepsilon [MR^{-1}] \text{ and } p_2 \ \varepsilon [MR^{-1}]$$

The dimension for income of the consumer is

$$Y \ \varepsilon \ [MT^{-1}]$$

When the commodity one is taken as numeraire commodity

$$Y = p_1 x_1 + p_2 x_2$$

$$\frac{Y}{p_1} = x_1 + \frac{p_2 x_2}{p_1}$$

Dimensionally three components of the equation are:

$$(1)\ Y / p_1 \varepsilon \frac{[MT^{-1}]}{[MR^{-1}]} = [MT^{-1} M^{-1} R_1] = [R_1 T^{-1}]$$

$$(2)\ p_2 x_2 \varepsilon \frac{[MR_2{}^{-1}][R_2 T^{-1}]}{[MR^{-1}]} = [MT^{-1} . M^{-1} R_1]$$

$$\varepsilon[R_1T^{-1}]$$

$$(3)\ x_1 \qquad \varepsilon[R_1T^{-1}]$$

This shows that dimensionally the above equation is homogeneous:

$$\frac{Y}{p_1} = x_1 + \frac{p_2x_2}{}$$

$$[RT^{-1}] = [RT^{-1}] + [RT^{-1}]$$

$$= [RT^{-1}]$$

Keynesian Theory of Employment

J.M. Keynes in "The General Theory of Employment Interest and Money" explained a theory of employment in which he distinguished three types of goods: capital goods, consumer goods and labour; and specified the national income equation:

$$Y = C+1$$

If the symbols of income (Y), consumption (C) and investment (I) are all in real terms the equation is dimensionally wrong because C and I are in two different dimensions and hence they are not additive.

When $Y = C + S$; and accordingly $S = 1$ is analysed the symbols are in the monetary units. Keynes effective demand analysis is in real terms hence unless the terms of assigned dimensions the equation cannot be consistent. In a developed economy unemployment is caused by the deficiency of effective demand, since there is no deficiency of savings for the purpose of investment. In real sense since investment means investment in fixed and variable capital of labour which are again in different dimensions the equation cannot be homogeneous unless everything is transferred to monetary units.

Schumpeter's Process of Money

J. M. Schumpeter assumed that investment is financed by bank-credit, when the innovater invests it, it increased money income and prices and helps to create a cumulative expansion throughout the economy. He provided a monetary equation:

Stock of money = (Price) (X : flow of consumer goods)

Dimensionally this means:

$$[M] = [MR^{-1}] [RT^{-1}]$$

$$= [MT^{-1}]$$

This shows that the equation is not dimensionally homogeneous. It can be made homogeneous by multiplying the right hand side of the equation with the velocity of circulation of money : V. The equation is not logical because a flow of money cannot be equal to the stock of money. Earlier we found that $V \varepsilon [T^{-1}]$ hence if V is multiplied to M the equation is homogeneous. $[MT^{-1}] = [MT^{-1}]$

Piogu's Theory of Employment

In the equilibrium of employment Piogu accepted the supply of labour as an abstract number or a constant in the following equation: W = w N d

Where W is the wages bill and w is the wage unit, and Nd quantity of employment.

Dimensionally

$Nd \; \varepsilon \; [R_a]$; Wage unit $w \; \varepsilon \; [MR_a^{-1}]$ and $w \; \varepsilon \; [MT^{-1}]$

Therefore $[MT^{-1}] = [MR_a^{-1}] [R_a]$

$$= [M]$$

This shows that the equation is not dimensionally homogeneous.

The main defect of this equation is that the supply of labour is taken as constant: which in fact is a flow. Keynes in his "General Theory" taken the supply of labour as a flow. The supply of labour can be derived from the above equation.

$$W/w = Nd$$

In dimensions:

$$\frac{[MT^{-1}]}{[MR_a^{-1}]} = [MT^{-1}][M^{-1}R_a]$$

$$[R_a T^{-1}] = [R_a T^{-1}]$$

This shows that supply is a flow, and when this is replaced in the above equation it becomes dimensionally homogeneous:

The fresh equation is $W = (w)\,(W/w)$

Dimensionally:

$$[MT^{-1}] = [MR_a^{-1}]\,[R_a T^{-1}]$$
$$= [MT^{-1}]$$

13

Meta Analysis

Review of studies made earlier is an important aspect of research methodology. Before finalising the research project a researcher has to go through the results of the earlier studies. In social sciences there are no specific methods for synthesising results. The traditional narrative method used for review, comparing and synthesising research results suffers from many defects.

Five important defects of the reviewers in the traditional narrative method has been pointed out by J.B. Jackson:

1. The reviewers generally do not critically examine the evidence, methods and conclusions. Mainly they take only the final results for comparison.
2. Reviewers frequently use crude and misleading representations of the study results.
3. Most of the reviewers focus their discussion and analysis partially, without taking the study as a whole, which is suitable or favourable to their own studies.
4. Reviewers fail to assess possible relationship between the characteristics and the final results of the studies.
5. Researchers who review the studies, report nothing or very little about the methods of reviewing. Therefore a reader is not in a position to judge the validity of the conclusion.

The seventies witnessed a large scale output of research studies in social and behavioural sciences and made informal narrative technique of research review and interpretation inoperative. A researcher become helpless to scan hundreds of research studies and assess them without any standard statistical methods. R.J. Light and P.V. Smith tried first to give serious attention to methodological and technical problems of research interpretation.

Types of Analysis

Analysis can be classified into three type on the basis of the source of data. Traditionally the two types of analysis are primary and secondary analysis, but now the third tier of analysis, synthesising research studies, is came to know as Meta Analysis.

Primary Analysis: Analysis of data collected from field survey is known as primary analysis. Primary data is raw data and its use in analysis provides true picture of relations in the study. Most of the research studies are primary analysis.

Secondary Analysis: When the data collected for a study through primary sources are used for another study it is known as secondary analysis. Here the old data is used to answer original research questions with better statistical techniques or used to answer new questions. Secondary analysis is also an important feature of research and evaluation.

Meta Analysis: Meta in Greek means 'over'; hence meta analysis means overall analysis. Contemporary research reviewing is more technical and statistical than narrative. The integration of research studies through statistical methods is now known as "meta-analysis". In meta analysis the studies already conducted in primary or secondary levels are taken for analysis. It is the analysis of the analyses. The essential character of meta analysis is that it is the statistical analysis of the summary findings of many empirical studies.

Development of Meta Analysis

Even though meta analysis, as a systematic method, is of recent origin, but some attempts can be traced back from 1931. Jay Lash for the first time used statistical methods in integrating the research results in agriculture, where he made an average of correlations of different findings. In 1946 George Snedecor made tests for the heterogeneity of a set of correlation coefficients. Later in 1949 Samuel Stouffer described a method of combining probabilities which is one of the important techniques of meta analysis to-day. Of course Karl Pearson (1933) and Ronald Fisher (1932 and 1933) developed procedures for combining probabilities, but by then meta analytic methods were not developed to use Pearson's formulas. F.M. Mostteller and R.R. Bush broadened the Fisher and Pearson's perspective and evolved several methods of combining independent probabilities available to social scientists. It has been the continuous efforts by Robert Rosenthal from the early sixties and G.V. Glass and his group in the seventies that established meta analysis as an important analysis for the social scientists.

Retrieving Research Results

The first tasks of a reviewer is the identification of the studies to be included for meta analysis. Generally the reviewers are not very much serious about locating the study and never bother to scan the vast literature. In a study J.B. Jackson revealed that about 8 per cent of the reviewers only searched bibliographies of previous reviews of the topic.

Source of empirical literature in the social sciences is divided into two classes: primary source and secondary source.

1. ***Primary Source:*** Archival periodical literature, i.e. journals, dissertations and theses and fugitive literature of government reports, papers from scholarly meetings, reports of public agencies are coming under primary source.

2. ***Secondary Source:*** Reviews in the periodicals and periodical reviews and abstracts are regarded as secondary source. The secondary source facilitates the identification of the primary source that will be taken for reviews. The technology of information storage is rapidly developing in this country to supply bibliographical and abstract information to the researchers.

There are four major classes of documents in which a reviewer can find studies relevant to his study.

1. Books including unauthored books, edited books and chapters in edited books.

2. Journals, including professional journals published newsletters, magazines and newspapers.
3. Theses, including doctoral, masters and bachelor's theses.
4. Unpublished works, including technical reports, grant proposals, convention papers, film cassette recordings etc.

Reliability of the Literature

Reliability of studies from the above sources poses a problem whether to accept all the sources or only few reliable sources. The main object of meta analysis is to provide an accurate, impartial, quantitative description of the findings in a population of studies on a particular topic. This may be done either from the existing population or from the sample representative. No survey is considered as valid if a sizable stratum of the population remains unrepresented in the sample.

One of the most important sub-set in research integration is the sub-set of unpublished studies. Most of the reviewers take only the published studies in books and journals and bother less to search for the unpublished materials. Most of the dissertations and theses remain unpublished hence remain aloof from the reviewers view. If a meta analysis contains only published studies it would wholly unrepresentative of the population. A study by Smith, Glass and Miller in 1980 of 11 meta analysis of experimental literature show that the average effect size is highest with 0.64 for journals while the effect size ranges between 0.48 and 0.58 for the theses and unpublished literature. Non inclusion of unpublished studies in a meta analysis may produce misleading generalisations.

Abstracting Results

After location of the study of the next task of a reviewer is to decide what information is to be abstracted from each document. The main information for the statistical analysis to be recorded should be the significance level and the size of the effect. Apart from these two, several other information have to be recorded for a correct assessment of the studies. Of course it depends on the nature of the study one undertakes, but the studies retrieved the following information are typically recorded by a researcher.

1. Complete reference for the purpose of bibliography.
2. Authors full name and address.
3. Sex of the data collector in the study.
4. Status of the data collector, i.e. faculty member, doctoral student, graduate student etc.
5. Sex of the subjects or respondents.
6. Relationship of the data collector to the author of the study.
7. Nature of the subject sample, i.e. where and how obtained.
8. Unusual design features if any.
9. Procedural controls for cheating and observer behaviour.
10. Moderating variables, variables associated with differences in obtained results, the direction of their effects size and significance level.

Reliability of Retrieval

The abstracts of the documents collected for meta analysis must be reliable. Reliability of locating research results and coding of study features is an important aspect of meta analysis. Two researchers may abstract different information from the same study for meta analysis and proportions of agreement on specific items coded by the two researchers can be calculated. On the basis of median proportions the reliability of a study document can be judged. The study with a higher median proportions is accepted as more reliable. Of course correlations of the two studies would get more confidence in the interpretation of data.

For most of the reliability assessment coding is essential for significance level and effect size estimates, but generally no reliability data are available for the estimation of significance level. Cooper and Rosenthal in a study in 1980 asked 19 meta analysts to decide whether a given set of 7 studies supported the rejection of null hypothesis. The interclass correlation with 19 × 7 data matrix showed a correlation of 0.969 which indicated a high degree of reliability.

Assessing Research Results

Error making is a normal phenomenon. The authors of studies might have made mistakes during the recording of data from the experiment or in the analysis. Meta analyst has to go through the original papers to find out the errors and correct them wherever possible: before taking up a study for meta analysis. When there is an error in recording in the original study it is very difficult to detect it. In most of the studies the error is defined as misrecording that was either seen or heard by the data recorder. In most of the studies the data is not provided in the useable form. Hence the meta analyst has to estimate error from the data available from the study. Of course small rate errors would not affect the study results.

Removal of all the errors from a study is probably not possible or even desirable from a cost/benefit perspective. But one should not take errors as universal and remain content with a study if it has errors. The main concept here is awareness of errors. This may generate more task attitude towards errors than what it is presently shared. It is believed that poor investigators make errors and good investigators don't make mistakes. A meta analyst's aim is not to find mistakes of others but to get the answers right.

Quality of Research Results

One of the major criticisms of meta analysis is that both poor and good studies are summarised in the analysis. But till now the traditional reviewers are also doing same thing. It is very difficult to give quality weights, and if at all it is given, one would assign high weights to studies whose results one favours and low for those not favoured.

A method has been evolved to assess quality of the studies. Each study has to be coded by excellent methodologists on a 9 point rating scale of the studies. Their quality assessment would be made twice as follows:

First, based only on their reading of the methods section: Second, again based on their reading of the methods plus the results section of the study.

On the basis of these ratings an "effective reliability" index can be calculated as per Spearman-Brown formula:

$$R = \frac{n\,r}{1 + (n-1)\,r}$$

Where R = Effective Reliability

n = Number of judges

r = Mean reliability of all n judges. If the number of judges are 5 then 10 correlation coefficients have to be computed and mean of these correlations have to be utilised. Hence for n judges n (n – 1)/2 correlations have to be estimated as per the weightage scale of the judges.

When the number of judges are small this formula may be useful but when large number of judges are taken a large number of correlation coefficients have to be computed which would be a time consuming process. Guilford has provided a simpler formula to calculate estimated R, which can be used for large number of judges.

$$\text{Rest} = \frac{\text{MS studies} - \text{MS residual}}{\text{MS studies}}$$

The mean reliability is given by

$$\text{rest} = \frac{\text{MS studies} - \text{MS residual}}{\text{MS studies} + (n-1)\ \text{MS residual}}$$

Table 13.1 shows the quality rating of three judges for five studies on a scale 1 to 7, and Table 13.2 summarises the calculations.

$$X_j^2 = 196 + 484 + 576 = 1256$$

$X_i^2 = 792$ $\quad X_j^2 = 1256$ $\quad T^2 = (60)^2 = 3600$

$C = T/ab = 3600/15 = 240$ $\quad a = 3$ (judges); $b = 5$ (studies)

Sum of Total Square $X_{ij}^2 = 282$

$ST = x_{ij}^2 - C = 282 - 240 = 42$

Squares of the A, B, C columns of Table 13.1

25	36	49
9	36	16
9	16	36
4	4	9
1	16	16
Total 48	+ 108	+ 126 = 282

SS (sum of squares) of judges: $(x_i^2 / a) - c = (792/3) - 240 = 24$

SS of studies: $(x_j^2 / b) - C = (1256/5) - 240 = 11.2$

Residual: $ST - SS_{jd} - SS_{st} = 42 = 24 - 11.2 = 6.8$

Table 13.1: Judges Rating of Research Quality

Studies	*Judges*				
	A	B	C	X_1	X_1^2
1	5	6	7	18	324
2	3	6	4	13	169
3	3	4	6	13	169
4	2	2	3	7	49
5	1	4	4	9	81
	14^2	22^2	24^2	60	792

Table 13.2: Summary of Calculations

Source	*SS*	*df*	*MS (mean square)*
Studies	24	4	6.00
Judges	11.2	2	5.60
Residual	6.8	8	0.85

On the basis of MS studies and MS residuals, R_{est} and r_{est} *can be calculated as under:*

For the example of three judges and five studies the estimated R can be calculated in the formula given earlier:

$$R_{est} = \frac{MS_{st} - MS_{res}}{MS_{st}}$$

$$= \frac{6 -- 0.85}{6} = 0.858$$

$$R_{est} = \frac{MS_{st} - MS_{res}}{MS_{st} + (3-1)MS_{res}}$$

$$= \frac{6 - 0.85}{6 + (2)0.85} = 0.669$$

Finding effective reliability through variance approach is slightly different from the correlation approach as per the Spearman-Brown formula. In the later formula the figures are slightly higher but negligible in comparison to variance approach which is less cumbersome for calculation.

Defining Study Results

In a study the most important result one wants finally is the relationship among the variables. Hence the two important information a meta analyst wants are:

(i) Estimation of the magnitude of relationship between the variables or the effect size.

(ii) An indication of the accuracy or reliability of estimated effect size. This may be known by the test of significance of the difference between the obtained effect size and the effect size expected under the null hypothesis that there is no relationship between the variables.

Results of a study with respect to any given relationship can be expressed as an estimate of effect size and a test of significance. A general relationship between the two can be expressed as follows:

Test of significance = (size of effect) × (size of study)

Pearson's r is regarded as most important effect size estimate. But Hedges, Cohen and Glass used different formulas for the estimation of effect size. Table 13.3 gives five different formulas for the computation of test of significance of a study.

In the equation numbers 1 and 2, ϕ is merely Pearson's r applied to dichotomous data; i.e. data coded as taking on as two values. In equation (4) M is the mean of the two compared groups and S is the unbiased estimate of population standard deviation. This is also termed as Hedges g.

Table 13.3: Relationship between Test of Significance and Effect Size

Equation	*Test of Significance*	=	*Size of effect*	×	*Size of study*
1	$X^2(1)$		ϕ^2		N
2	Z		ϕ		$\sqrt{N}$
3	t		$\sqrt{\frac{r}{1-r^2}}$		$\sqrt{df}$
4	t		$(M_1-M_2)/S$		$\frac{1}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$
5	t		$(M_1-M_2)/6$		$\sqrt{\frac{n_1 n_2}{n_1+n_2}}$

In the equation (5) σ is the pooled sum of squares divided by N. This formula of effect size is termed by Cohen as d. Glass also uses $(M_1 - M_2)/S$ as effect size estimation, but S here as the standard deviation of the control group. S is computed as:

$$S = \frac{(X-\overline{X})^2}{(n_c-1)} \cdot \frac{1}{2}$$

Most of the studies generally do not provide effect size estimates but only the test of significance. Hence a meta analyst has to compute his own effect size from the tests of the significance provided. Accordingly r, ϕ, and d can be computed through these formulas:

$$\text{Eq. (6)}: \phi = \sqrt{\frac{X^2(1)}{N}}$$

$$\text{Eq. (7)}: r = \sqrt{\frac{t^2}{t^2 + df}}$$

If non of these tests of significance are found in the study still one can estimate effect size from the p levels. First p levels has to be converted into Z values and r can be computed as per the following formula:

$$\text{Eq. (8)}: r = \sqrt{\frac{Z^2}{N}}$$

$$= \frac{Z}{\sqrt{N}}$$

All these equations produce product moment correlation coefficients. It makes no difference whether they are in dichotomous, continuous or in rank form. Even when Cohen's d is available it can be converted as r as per the formula:

$$\text{Eq. (9)}: r = \frac{d}{\sqrt{d^2 + 1/pq}}$$

Here in the equation p is the proportion of the total population that is in the first of the two groups and q = 1 – p.

Adjusting Effect Size Estimates

When large number of studies are reviewed and r values are taken, r's sample from that population becomes more and more skewed. Hence Fisher advised that assume r distribution as near normal and transform r into Z_r for all purposes of analysis. So instead of r the value of Z_r is generally used as good estimate of effect size. The transformation formula is:

$$\text{Eq. (10)}: \frac{1}{2} \log_e \frac{1+r}{1-r} = Z_r$$

Example: Suppose that an experiment provides these information: $n_1 = 4$; $n_2 = 8$; and hence $t(10) = 2.67$

$$\text{First compute } r = \sqrt{\frac{(2.67)^2}{(2.67)^2 + 10}}$$

$$= .658 = Z_r = .789$$

There is a small and negligible bias in Z_r. When N is very small and r population is very substantial the bias is more in Z_r. For all practical purposes it can be neglected. The bias can be removed from Z_r by calculating it in the formula of Fisher's.

Eq. (11) Estimated Bias = $\frac{r}{2(N-1)}$

For the example above bias can be calculated as:

$$\text{Estimated bias} = \frac{0.658}{2(12-1)}$$

$$= 0.30$$

This bias has to be removed from the obtained Z_r of 0.789 so the correct Z_r is

$$0.789 - .30 = 0.758$$

When Final Results are not Available

In a study when original data are not available to estimate effect size, but only a set of results available in p levels one can calculate mean level of significance and mean effect size for review.

A mean level of significance can be calculated in the following procedure. First convert the p levels into the standard normal deviates or Z and then find the mean of Z. This value can be converted to p value as a mean of significance level.

Suppose there are three p values of .25; .10; .001

Corresponding Z values are .67, 1.28, 3.09

Mean of Z values: $(.67 + 1.28 + 3.09)/3 = 1.67$

Converting this into p value is 0.046. This is the value of significance level which has to be accepted as the estimated level of significance for the study.

If there are several effect sizes in a study their simple mean can be calculated for the purpose of review. Instead of r if the effect sizes are available in the form of Cohen's d, Hedges g or in Glass's S its mean also can be calculated. Normally the effect sizes are given in the form of r; to find out the mean first transform each r into Z_r and then calculate the simple mean. If the r is also not there in the study and only Z values are available compute r value for each Z value when N, the number of sample units are known. The formula for r is

$$r = \frac{Z}{\sqrt{N}}$$

For the example given earlier the mean effect size can be calculated as under:

p levels :	.25	.10	.001
r :	.067	.128	.309
Z_r :	.07	.13	.32
Z_r :	.17	in which N = 100	

Hence the mean Z_r is the effect size of the study for the purpose of review.

Meta Analysis of Two Studies

The level of significance can be compared between two studies in order to know whether there is any significant difference between the two studies. For the calculation of significance testing first find out the p values recorded as one tailed, transfer it to corresponding Z values and then find out the difference according to the following formula:

Z distributed as = $= \frac{Z_1 - Z_2}{\sqrt{2}}$

Example:

p values (one tailed)	.04	.000025
Z value for above	1.75	4.06

Hence difference Z $= \frac{(4.06)-(1.75)}{1.41} = 1.64$

p value for Z = 1.64 is .05 one tailed. This shows that the two values differ significantly. But one should remember that one Z greater than another does not tell any information whether the Z was greater because of

1. the size of effect was greater
2. the size of study, i.e. N the sample units was greater or
3. both are greater.

Combining Studies

When significance test has to be computed for two studies the Z values are added together according to Stouffer Method. In this method first the p levels for each of the two studies are obtained and then the corresponding Z values are found. And finally combined Z is found in the following formula:

$$\text{Combined } Z = \frac{Z_1 + Z_2}{\sqrt{2}}$$

Each Z can be given weightage by its df, its estimated quality or any other desired weight. Thus

$$\text{Weighter } Z = \frac{w_1 Z_1 + w_2 Z_2}{\sqrt{w_1^2 + w_2^2}}$$

Example:

p values: .05 .0000001 one tailed in opposite directions.

Z value : –1.64 5.20

$$\text{Combined } Z = \frac{(-1.64) + (5.20)}{1.41} = 2.52$$

p value associated with Z of 2.52 is .006 one tailed or .012 two tailed.

Effect Size Estimation

In meta analysis the effect sizes of two studies can be compared or combined as in case of level of significance but with slightly different formula. Let us first take the comparison of two effect size estimates. Here one wants to know whether the results are reasonably consistent with each other or whether they are significantly heterogeneous. Generally r is taken as the index of effect size. First r's of the two studies are converted into Z_r and then the difference is found in the following formula:

$$\text{Difference in effect size} = \frac{Z_{r1} - Z_{r2}}{\sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}}$$

Example: Effect Size r : 0.60 –0.20

N : 15 100

Fisher's Z : 0.69 –0.20

$$\text{Hence the difference is :} \quad \frac{(0.69 - (-0.20))}{\sqrt{\frac{1}{12} + \frac{1}{97}}}$$

$$= 2.91$$

This is the Z of difference between the two effect sizes. The p value associated with it is 0.002 one tailed. This shows that the effect sizes differ significantly between the two studies.

Combination of effect sizes: Combination of effect sizes of two studies to get a representative figure calculate the simple arithmatic mean. But convert the r's into Fisher's Z_r before finding out the mean. The simple formula for mean Z_r or $\bar{Z}_r$ is

$$\bar{Z}_r = \frac{Z_{r1} + Z_{r2}}{2}$$

Then transform the mean $\bar{Z}_r$ into r from the table of Z_r. Of course the mean Z_r can be converted in the formula:

$$r = \frac{(e^2 Z_r - 1)}{(e^2 Z_r + 1)}.$$

Weighted mean of Z_r can also be found by weighing each Z_r by its df i.e. N – 3. The weighted formula is:

Weighted mean $\overline{Z}_r = \frac{w_1 Z_{r1} + w_2 Z_{r2}}{w_1 + w_2}$

Example:

Value of r : .80 –.30

Z_r 1.10 –1.31

$$\overline{Z}_r = \frac{(1.10) + (-0.31)}{2}$$

$= 0.395$

The Z_r of .395 is associated with r of .38

Voting Method of Integration

The most commonly used method of integrating research studies is known as Voting Method. Light and Smith characterised this method in a simple way. In this method, first the relationship between dependent and independent variables are established and the results are distributed into three categories:

1. Relationship significantly positive.
2. Relationship significantly negative.
3. No significant relation in either direction.

The maximum number of studies falling into a category is declared winner. This winner category is accepted as the true representative of the reviewed studies. This model of categorisation is assumed as the best estimation of the direction of the true relationship.

Even though this method is a simple one but it suffers from certain deficiencies. First, the method disregards the sample size. Large samples more frequently produce statistically significant findings than the small samples but in this method no consideration is given for the size of the sample. Another deficiency of the voting method is that it disregards good descriptive information. The tallies of statistical significance or insignificance tell little about the strength or importance of relationship which is vital information.

Counting Method

X^2 statistic can be utilised in comparing number of studies expected to reach a given level of significance with the number actually reaching that level. In this method a critical p level is set at .05; .10; or .01 and then the studies are classified into two counts: those who are reaching it in one and not reaching in the second. On the basis of information $X^2(1)$ statistic is calculated in the following formula:

$$X^2(1) = \frac{(O-E)^2}{E}$$

Where O denoted for obtained and E for expected number out of the total number of studies.

Example:

Count	*Studies reaching p = .05*	*Studies not reaching p = .05*	*Summation*
Obtained	12	108	120
Expected	6	114	120

The expected count is calculated on the basis of critical p level. In the above example the critical p level is .05, hence the expected number studies reaching .05 should be 120 × .05 = 6; accordingly the studies not reaching .05 is .95 hence the expected number of studies not reaching the critical level is 120 × .95 = 114.

$$X^2(1) = \frac{(12-6)^2}{6} + \frac{(108-114)^2}{114}$$

$$= 6.32$$

$$p = 0.012$$

Meta Analysis for Many Number of Studies

When number of studies are more than two or in large numbers there are certain problems in analysing them. When large number of studies are taken up for review there are two types of tests: Diffused tests and Focussed tests. In a diffused test one knows only whether they differ significantly or not but it provides no information how they differ or whether they differ according to any systematic basis. In a focussed test one knows whether the studies differ significantly among themselves in a theoretically predictable or meaningful way.

Comparison with Diffused Tests

Significance Testing: When more than two studies have to be compared, first their p levels have to be converted into standard normal deviat, Z. The statistical significance of the heterogeneity of the Z can be obtained from X^2 computed as follows:

$\Sigma(Z_j - \bar{Z})^2$ is distributed as X^2 with K–1 df.

Example:

p values :	.15	.05	.01	.001
Z values :	1.04	1.64	–2.33	3.090

$$\Sigma(Z_j - \bar{Z})^2 = (1.04 - 0.86)^2 + (1.64 - 0.86)^2 + (-2.33 - 0.86)^2 + (3.09 - 0.86)^2$$

$$= 15.79$$

This X^2 value with 3 df is significant at p = 0013. Hence the 4 p values compared are clearly significantly heterogeneous.

Effect Size Estimation: The main object of comparison here is to assess the statistical heterogeneity of three or more effect size estimates. Here r is taken as effect size estimate. First r's are converted into Fisher's Z_r and then the mean of Z_r is calculated according to the formula given below:

$$\bar{Z}_r = \frac{\Sigma(N_j - 3)\,Z_{rj}}{\Sigma(N_j - 3)}$$

Then X^2 value is found on the basis of the formula

$\Sigma(N_j - 3)(Z_{rj} - \bar{Z}_r)^2$ is distributed as X^2 with K–1 df.

Example:

r values :	70	.45	.10	−.15
N :	30	45	20	25
Z_r values :	87	.48	.10	−.15
N − 3 :	27	42	17	22

$$\overline{Z}_r = \frac{27(.87)+42(.48)+17(.10)+22(-.15).}{27+42+17+22}$$

$$= \frac{42.05}{108} = .39$$

For the estimation of X^2 value is then

$$X^2 = 27(.87-.37)^2 + 42(.48-.39)^2 + 17(.10+.39)^2 + 22(-.15-.39)^2$$
$$= 14.41$$

This value for K−1 = 3 df is significant at p = .0024. This indicates that the four values are significantly heterogeneous.

Comparing Studies with Focussed Tests

In a diffused test all the individual studies are taken on same footing. The size of the samples are also not known in these tests. For instance, even though all the four studies are analysing teacher-student relationship and tries to find out teacher expectancy effect; the age of the students might have different in the studies or the number of hours the teacher engaged with students might be different and accordingly specific relationship between the teachers and students is considered important. In a focussed test one learns how the studies differ themselves on the basis of different characteristics of the independent variables which can be theoretically predictable or predictable in a meaningful way. Given a set of p values for the studies of teacher expectancy effects one might want to know whether results from younger children show greater degree of statistical significance than the result from older children.

Significance Testing: The procedure followed for significance testing under focussed testing is as follows: Convert all p levels of the studies into standard normal deviate Z and compute the mean of Z values on the basis of contrast weight as per the formula given below:

$$\frac{\Sigma \lambda_j Z_j}{\sqrt{\Sigma \lambda_j^2}} \text{ is distributed as Z}$$

λi is the theoretically derived prediction of contrast weight for any one study, chosen such that the sum of the λ*i* will be zero, and Z_j is the Z for any one study.

Example:

One tailed p values :	.0000001	.0001	.21	.007
Z values :	5.20	3.72	.81	2.45
Hours of tutor-student contact per month	8	6	4	2
Weights of linear contrast for the study λ	3	1	–1	–3 = 0

$$Z = \frac{(3)5.20 + (1)3.72 + (-1)81 + (-3)2.45}{\sqrt{(3)^2 + (1)^2 + (-1)^2 + (-3)^2}}$$

$$= \frac{11.16}{\sqrt{20}} = 2.50$$

The Z value of 2.50 is significant at p level of .006 one tailed. This result shows that the four p values tend to grow linearly more significant as the number of hours of contact increases.

Effect Size Estimates: In comparing effect sizes of more than two studies one wants to ask more focussed question about the relationship. With the reference to the above example of tutor-student relationship, one wants to know whether these effects are increasing or decreasing linearly with number of hours of contact per month.

Computation: For the each effect size r, first compute its associated Z_r, and N–3; where N is the number of sampling unit on which each r is based. The common Z is computed in the following formula.

$$\frac{\sum \lambda Z_{rj}}{\sqrt{\sum \frac{\lambda_j^2}{w_j}}} \text{ is distributed as Z}$$

λ_j is the contrast weight determined from some theory for any one study, chosen such that the sum of the λ_j will be zero, w_j is the inverse of the effect size for each study. For the Fisher's Z_r transformations of the effect size r the variance is $1/(N_j-3)$ so $w_j = N_j-3$

Example:

Effect size r	:	.89	.76	.23	.59
Fisher's Z_r	:	1.42	1.00	.23	.68
Hours of contact per month	:	8	6	4	2
Contrast weight	:	3	1	–1	–3

$$Z = \frac{(3)1.42 + (1)1.00 + (-1).23 + (-3).68}{\sqrt{\frac{(3)^2}{9} + \frac{(1)^2}{9} + \frac{(-1)^2}{9} + \frac{(-3)^2}{9}}}$$

$$= \frac{2.99}{\sqrt{2.222}}$$

$$= 2.01$$

This Z value is significant at p = .002 one tailed. This shows that the four effect sizes tend to grow linearly larger as the number of hours of contact time increases. Interpretation of this relation must be very cautious. It should not be interpreted as giving a strong evidence for a causal relationship. But these moderator relationships are very valuable in suggesting the possibility of causal relationships.

Combining Studies with Focussed Tests

Significance Testing: Combining studies in the focussed tests on the basis of p levels for any number of results for three or more studies requires different method of computation than in diffused tests. The procedure is, for each p level the Z values are obtained which should be given in one tailed. The sums of the Z divided by the square root of the number K of the studies yields a new statistic distributed as Z.

$$\frac{\sum Z_j}{\sqrt{K}} \text{ is distributed as Z}$$

The studies can be given weightages on the basis of their df, its estimated quality or any other desired weight. The weighted Z can be computed as under:

$$\text{Weighted Z} = \frac{\sum w_j Z_j}{\sqrt{\sum w_j{}^2}}$$

Example:

One tailed p values	:	.15	.05	–.01	.001
Z values	:	1.04	1.64	–2.33	3.09

$$Z = \frac{1.04 + 1.64 + (-2.33) + 3.09}{\sqrt{4}}$$

$$= 1.72$$

The new Z value is associated with p value of .043 one tailed or .086 two tailed. One tailed p value is normally applied if the bulk of the findings are correctly predicted. These four studies can be given weightages on any theoretical grounds and can be computed as per the weighted formula given above.

Effect Size Estimation: Combining effect sizes is an easy method when the results of three or more than three studies are combined. Here one has to find out only mean of Z_r for the number of studies combined K. Weightages can be given to each study on the basis of its df, i.e. N–3 by its estimated research quality or by any other weights before data is finalised. The unweighted and weighted formulas are given below:

$$\bar{Z}_r = \frac{\sum Z_r}{K}$$

$$\bar{Z}_r = \frac{\sum w_j Z_{rj}}{\sum w_j}$$

Example:

Effect size r	:	.70	.45	.10	−.15
Fisher's Z_r	:	.87	.48	.10	−.15

$$\overline{Z}_r = \frac{(.87) + (.48) + (.10) + (-.15)}{4}$$

$$= .32$$

This Z of .32 is equal to corresponding r of .31.

Example:

By providing specific weightages the above example can be worked out again to find out the combined effect of the four studies. Weightages given on the basis of some theoretical grounds for the four studies are: 1.7; 1.6; 3.1; 2.5.

$$\text{Weighted } Z_r = \frac{(1.7)(.87) + (1.6) + (.48) + (3.1)(.10) + (2.5)(-.15)}{1.7 + 1.6 + 3.1 + 2.5}$$

$$= \frac{2.182}{8.90} = .24$$

This Fisher's Z_r corresponds to an r of .24.

Critics of the Analysis

Meta analysis is an attitude of data analysis applied to quantitative summaries of individual studies. It applies statistical methods to the task. Hence it is not a technique itself, it uses many techniques, measurements and statistical analysis to integrate research studies. But some have criticised the method in the following lines.

1. ***The meta analysis method mixes apples and oranges:*** One of the traditional criticism of meta analysis is that it approaches to integrate research studies is similar to mixing of apples and oranges which are generally unmixable. The contention is that it makes no sense in integrating findings of heterogeneous studies and aggregation of different sizes of studies are absurd. Only studies that are same in certain respects can be aggregated. But this criticism seems to be self-contradictory. If the studies

are the same in all respects there is no need to compare them since they would be obviously have the same findings within statistical errors. The only studies which need to be compared are different studies. Furthermore, this criticism seems to be baseless, because traditionally the researchers have been comparing and synthesising the studies for conclusion on a 'literary plane' adopting same method, that is, mixing the apples and oranges. In comparison to the traditional method meta analysis is made more objective and quantitative for better analysis.

2. ***Meta analysis use data for poor studies:*** Some critics believe that the meta analysis approach 'advocates low standard of judgement' of the quality of the studies undertaken. In this criticism it seems the critics are confused between 'research design' and 'research findings'. A study can be judged as good or poor on the basis of quality of research design. Research design has a logic of its own, but its logic not appropriate to research integration. Generally when a research study fails to meet conditions of which the critic is fond of, the study is eliminated from review considering it as poor study.

A meta analysis of poor studies, i.e. poor in some portion of research design may provide a good result when integrated. Meta analysis also records methodological weaknesses in the original studies and examines their co-variance of study findings. The influence of study quality on findings has been regarded consistently as an empirical, a posteriori question, not an apriori matter for exclusion of large number of studies from consideration. Inclusion of poor studies over a large sample provides good results and do not effect the overall conclusion to be drawn from the review.

R.K. Yin and others prescribed three criteria for the judging the quality of a study. They are

(i) Presence of operational measures of innovative device and outcome.

(ii) Presence of a relevant research design.

(iii) Overall adequacy of evidence in relation to conclusion.

Based on these characteristics of quality meta analysis takes due consideration for quality. Several meta analyses conducted by prominent meta analysts in recent years reveal that the results are more consistent when the sample size is widened inclusion of all type of studies irrespective of whether the study is poor or good.

3. ***Selection bias in recorded research:*** Meta analysis is dependent on the findings that researchers report. Its findings will be biased if there are systematic differences among the results of research that appear in journals versus books versus theses versus unpublished papers.

Study by different meta analysts show that the journals have more favourable results than books, and the published books have more favourable results than either theses or unpublished works. Meta analysis includes all the five sources for review including theses and unpublished works. In traditional research review the researchers take only books and papers published in journals, and in fact more biased than the meta analysts.

4. Lumpy or non-independent data. It is alleged that meta analyses are conducted on large data set in which multiple results are derived on the same study; this renders the data non-independent and gives one a mistaken impression of the reliability of the results.

The principal implication of the non-independence is a reduction in the reliability of estimation of averages or of regression equations. A facile solution for the non-independence would be to average all the findings in a study before it is taken up for meta analysis. In one study if there are 4 variables each with .2 effect size and in a second study there are 3 variables with effect size of .6 each then only the effect sizes .2 and .6 have to be taken for meta analysis but not the total of seven variables in two studies. Hence the representative figures have to be finalised first at the study level, by averaging them, for better accuracy of estimation. If this method is adopted meta analysis cannot be criticised as lumpy.

Results of a meta analysis truely represent the studies reviewed or not, is not dependent on the techniques of meta analysis but on the nature of the research design and what actually revealed in the research study. Most of the research papers and books are defective in the sense that the detailed information is not revealed in the studies, and the meta analysis based on the revealed information makes it inaccurate which cannot said that the meta analysis is defective. Most of the research studies are defective on these grounds:

1. The important conditions that mediate the findings of the research studies are not generally reported in the written accounts of the research. Hence without getting a full picture of the research design and the conditions influencing the results, a meta analyst is unable to take the real indicators for analysis.
2. The social and behavioral sciences are at very primitive stages in accounting for their own findings. Either the details are not given because they are more complicated or the researcher prefers to keep it for himself, for the reason known to him; in either case the meta analyst is helpless in computing the true picture of the overall indicators in meta analysis of the studies.

The Strengths of Meta Analysis

In recent years meta analysis is gaining popularity among the researchers because of its strength in objective analysis. They are

(a) Quantitative review has been viewed as an efficient way to summarise large research literature. It is possible to reach stronger results in comparison to impressionistic literary reviews.

(b) Meta analysis takes both good and poor studies on the same footing hence it is more objective than the traditionally followed literary review.

(c) Meta analysis is helpful in highlighting the gaps in the literature and provides incentives into new directions for research. As a result of the development of the meta analysis the researchers are now more conscious about their research design and try to give more detail accounts of their research in arriving the final conclusions.

Bibliography

Ackoff, Russel L., (1961) *The Design of Social Research,* University Chicago Press.

Aldenderter, M.S. & Blashfield, R.K., (1984) *Cluster Analysis*, Sage Publications.

Allen, T. Harrdl, (1978) *New Methods in Social Science Research,* Praeger Publishers, New York.

Anderson, T.W., (1958) *Introduction to Multivariate Statistical Analysis,* John Wiley & Sons, New York.

Bhatnagar, S.K., (1986) *Network Analysis Technique,* Wiley Eastern Ltd.

Boulding K.E., (1966) *Economic Analysis* Vol. I: Micro-Economics, 4th Ed. Harner & Row, New York.

Cattele, Raymond B., (1952) *Factor Analysis for Psychologists and Social Scientists,* Harper and Row.

Chou Ya Lun, (1969) *Statistical Analysis,* Holt Rinehart & Winston Inc., New York.

Cromwell, Joff B., (1985) *Univariate Tests for Time Series Model,* Sage Publications.

Deming, W. Edward, (1960) *Sample Design in Business Research,* John Wiley & Sons.

Deo, Narsing, (1986) *Graph Theory,* Prentice Hall.

Dunteman, G.M., (1984) *Introduction to Multivariate Analysis,* Sage.

Everitt, Brain, (1974) *Cluster Analysis,* Hineman Educational Books, London.

Ferber, R., Verdoorn P.J., (1962) *Research Methods in Economics And Business,* Macmillan & Co.

Festinger, L. Katz, Daniel Edt. (1976) *Research Methods in Behavioral Science,* American Publishing Co.

Freedman P., (1960) *The Principles of Scientific Research,* Pergmon Press, New York.

Glass, Gene V., et al., (1981) *Meta Analysis in Social Research,* Sage Publications.

Gortes, Fernando, et al., (1974) *Systems Analysis for Social Scientists,* John Wiley, New York.

Goode, William J., & Hatt, Paul K., (1952) *Methods in Social Research,* McGraw Hill, New York.

Hararay, Frank, (1985) *Graph Theory,* Narora Publishers.

Hurman. H.H., (1960) *Modern Factor Analysis,* Chicago Press.

Hildebrand, Devid K, et. al (1977) *Analysis of Ordinal Data,* Sage Publishers.

Jong, Frits J. De, (1987) *Dimensional Analysis for Economists,* North Halland.

Kim, Jae-on, Muller, Charles W., (1978) *Introduction to Factor Analysis*, Sage Publication.

Kirk, Jerome, Miller, & Mare L., (1986) *Reliability and Validity in Qualitative Research,* Sage Publications.

Klecka, William R., (1983) *Discriminant Analysis,* Sage Publications.

Knoke, David, Kuklinski & James H., *Network Analysis,* Sage.

Lachenbruch, Peter A., (1975) *Discriminant Analysis*, Habner Press, New York.

Laster, Albert (1982) *Project Planning and Control,* Butterworth.

Levine, Mark S., (1994) *Canonical Analysis and Factor Comparisons,* Sage Publications.

Lewley, D.N., & Maxwell, A.E., (1971) *Factor Analysis as Statistical Method,* ButterWorths, London.

Lews-Beak Michael S., (1980) *Applied Regression: An Introduction,* Sage Publication.

Morgen, David L., (1988) *Focus Group as Qualitative Research,* Sage Publications.

Ostrom, Charles W., (1990) *Time Series Analysis: Regression Technique,* Sage Publications.

Optner, Stanford L., (1988) *Systems Analysis for Business Management,* Prentice Hall, New Delhi.

Robbins, L., (1983) *Nature and Significance of Economic Science,* Macmillan, London.

Rosenthal, Robert (1984) *Meta Analytic Procedures for Social Research,* Sage Publications.

Sayrs, Lois W., (1984) *Pooled Time Series Analysis,* Sage Publications.

Schumpeter, J.A., (1957) *History of Economic Analysis,* Oxford University Press.

Sharma, A.V.B., et. al. (1983) *Research Methods in Social Sciences,* Sterling Publishers, New Delhi.

Sharma, Motilal Ed. (1985) *Systems Approach. Its Application in Education,* Himalaya Publishing House, Bombay.

Shrivastava, U.K., et. al. (1983) *Quantitative Techniques for Managerial Decision Making,* Wiley Eastern Ltd., New Delhi.

Whitehouse, Gary E., (1973) *Systems Analysis and Design using Network Technique,* Prentice Hall.

Wolfe, Hanry Dean, (1966) *Business Forecasting Methods,* Molt, Rinehart & Winston, New York.

Yamane, T., (1980) Statistics: *An Introductory Analysis,* Harper & Row.

Index

A

Activity analysis, 217-234
  allocation of resources, 229-230
    availability of monetary resources, 230
    human resources, 230
    machinery, 230
    material resources, 230
  control and monitoring, 234
  cost analysis, 231-232
  crashing of projects, 232-234
    cost of crashing, 232-234
  critical path method, 217-218
  designation of activities, 220-221
  drawing of networks, 219-220
  dummy activities, 221
  estimation of time, 224
  event time, 224-225
  finding the critical path, 226-228
  network planning, 218-219
  – rules, 221-223
  PERT and CPM, 223-224
  probabilities of completion, 228-229
  programme evaluation and review technique, 218
  resource scheduling, 230-231
  slack time, 225-226
Adjacent matrix, 210
Adjusted Correlation Matrix, 110, 111, 113
Administrative system, 164
Advocates low standard of judgement, 281
Allsis, Maurice, 235
Alpha factoring, 109
Alternative hypothesis, 52, 53, 55
Analyses in research, 1-9
  analysis, 6-8
    marginalism, 7
  analytical research, 1
  applied research, 2
  collection of data, 5-6
  definition of the problem, 4
  descriptive research, 1
  equilibrium, 7-8
  fundamental research, 2
  interpretation and report writing, 8-9
  laboratory research, 2
  research design, 4-5
  – methods, 2-4
  review of literature, 4
Analysing consumer behaviour, 137
Analysis of nominal and cardinal data, 8
Analysis of nominal data, 70-84
  association for IxJ tables, 82-83
  chi-square test, 73-74
  comparison of observed and expected values, 76-77
  components of chi-square, 77

  - contingence coefficients. 83
  - dependent and independent variables. 72
  - drinking habit of people. 71
  - expected value, 74-75
  - measures of association. 79-82
    - correlation coefficient. 82
    - cross product ratio. 80-81
    - measures based on chijsquare. 81-82
    - – in 2x2 tables. 79
    - percentage method. 79-80
  - notations, 72-73
  - null hypothesis. 74
  - number of categories. 71-72
  - partitioning chi square. 77-79
  - relationship between the variables. 77-79
  - sampling, 72
  - tichuprow's T value. 83-84
- Analysis of ordinal data. 85-94
  - Cohen's Kappa. 90-91
  - expected error. 88-89
  - modified measure. 94
  - observed error. 88
  - ordinal association, 86-89
  - probability measure, 89-90
  - quadrant measure. 92-94
  - statistical independence. 90
- Arthmatic mean. 156
- *Aswa* (horse), 136
- Atomic theory of matter. 138
- Auto-correlation, 35. 50
- Automatic control, 175

B

- Bachhayana, Muni. 136
- Beauty, 136
- Bibby. 52
- Binary data, 142
- Biological and social sciences. 136
- Bivariate correlation coefficient, 66
- Bivariate regression, 40. 63
- Black, 200
- Boulding, K.E.. 7
- Bridgeman. P.W.. 235
- Brisha (bull). 136
- British scientists. 161
- Burt, R.S., 203
- Bush, R.R., 260
- Business cycle. 18. 19
- Business government policy, 162
- Businessmen. 24

C

- Capital goods, 255
- Capitalist economy. 182
- Cattell, R.B., 115
- Celcius, 40
- Central bank, 7
- Central Bank Policy. 243
- Central tendency. 102
- Chi square, 117
  - statistic, 134
- Chitrini, 136
- Christianity, 69
- Classical algebra, 238
- Classification of cases into groups, 123-123
- Cliques, 203
- Cluster analysis steps in cluster analysis: groupings, 138
- Cluster analysis. 8. 136-160
  - clumping technique. 149
  - cluster tests. 160
  - concept of similarity, 139-140
    - distinguishability of non-identicals, 140
    - indistinguishability of identicals, 140

        symmetry. 139
        triangular inequality. 140
    density or mode seeking technique. 149
    distance measures. 146-148
        Euclidean matric measure. 146-148
    divisive method. 154-158
        monotonic division, 154-158
    hierarchical technique. 148-149
        agglomerative method. 149
    optimisation partitioning technique. 149
    quantitative data. 145-146
    selection of clusters. 159
    similarity coefficients. 140-144
        grower general similarity coefficient. 144
        Jaccard coefficient. 142-144
        simple matching coefficients. 141-142
    similarity measures. 138-139
    steps in cluster analysis. 138
    – – computation of similarities. 138
    – – definition of variables. 138
    – – selection. 138
    techniques of clustering. 148-153
    ward method of clustering. 153-154
Cluster sampling. 5
Cohen. 268. 269
Command and control. 162
Communality. 107
Communilism. 182
Complete linkage method. 149
Consumer goods. 186. 255
Consumption of goods and services. 7
Contingency coefficient. 83
Control. 175. 218
Cooper. 262
Correlation coefficient. 82
Correlation matrix. 109. 117. 118. 127. 128
Correlation of coefficients. 104
Cosign theta. 105
Co-variance matrix. 127
CPR. 80. 81
Critical path method. 216. 226
Cross product matrix. 128. 134. 135. 217. 218
Cybernetics or Control and Communication in the Animal and the Machine. 175
Cybernetics. 175
Czekanowski. J.. 136

D

Danglers. 223
Das Capital. 182. 252
Demand. 238
Dendro. 148
Dependent variable. 48. 49. 82. 80. 93. 97. 103. 273
Descriptive research. 1
Dichotomised data. 72
Dichotomous variables. 79
Dimension. 236
Dimensional analysis. 8. 235-257
    Buckingham theorem. 250-252
    dimensional analysis as an aid to memory. 249-250
    – – of some economic theories. 252-257
    – constants. 246-248
    – entities. 245-249
    function of measurement. 235-236

fundamental propositions. 240-241
Keynesian theory of employment. 255
Marxian theory of production. 252-255
objectives of dimensional analysis. 244-245
Piogu theory of employment. 256-257
primary dimensions. 237
Schumpeter process of money. 255-256
secondary dimensions. 237-241
system of measurement. 235
type of equations. 241-244
    behavioural equations. 242-243
    institutional equations. 243
    technical equations. 244-245
units of measurement. 236-237
Discrimating variables. 133
Discrimation functions. 133. 135
Discriminant analysis. 8. 122-135
Discriminant variables
    between groups matrix. 128-129
    centroids. 128
    classification of cases. 133
    classification with discriminant functions. 134-135
    derivation of co-efficients, 126
    discriminant function. 125
    – scores, 129-131
    generalised distance functions. 134
    limitations, 124
    main use of the analysis. 123-124
    mathematical model, 124-125
        cases. 124-125
        discriminant variables, 125
        groups. 124
    significant functions, 131-133
    simple classification function. 133-134
    total matrix. 126-127
    within group matrix. 127-128
Divisive method. 149
Doctoral student. 262
Doctors, 122
Dummy variable. 69. 201

E

Economic analysis. 7
Economic Regulation Process and Technique. 181
Economic thought. 7
Economics and management science. 11
Economics. 7. 137. 164
Economists. 24
Economy. 182
Edge, 188
Effective reliability, 263
Eigen value. 115. 116. 129, 131, 132
England, 161
Equilibria, 175
ESD, 47
ESS. 154
Euclidean, 155
Events, 220
Eviritt, B., 137
Expected error. 88, 89, 90
Experimental psychologists, 85
Explanatory model, 46
Explanatory variable, 11
Exploratory factor analysis, 117
Exponential trend line, 29, 31
Export and import of goods, 7

Export factor, 1

F

Factor analysis. 8. 95-121
  adjusted correlation matrix, 110-113
  aims of factor analysis, 97
  assumptions
    linearity, 100-101
    relationship of unique factors, 101
    standardisation of variables, 101
  characteristics of the variables, 98-100
  common factors, 99-100
  confirmatory analysis, 98
  correlation matrix, 106
  covariance and correlation, 102-103
  degree of explanation, 113-114
  error term, 100
  exploratory analysis, 98
  extraction of factors, 109-110
  general principle of retention, 115
  limitations, 118-119
  main characteristics, 96
  method of least squares, 116-117
  nature of co-variance, 103-104
  number of underlying factors, 117-118
  oblique model, 107
  retention of factors, 114-115
  rotation of factors, 119-121
  scientific research, 96-97
  screen test, 115-116
  structure matrix, 107-109
  types of analysis, 104-106
    factor loadings, 105-106
    oblique model, 104-105
    orthogonal model, 104
  types of factor analysis, 98
  unique factors, 100
Factor analytic techniques, 149
Factor loading, 105, 106, 113, 115, 119, 120
Factorial litter, 116
Farenheit, 40
Feedback path, 171, 178
Feudal economy, 182
Field-setting research, 2
First star, 211
Fisher, R.A., 122, 260
Forward path, 171
Fourier, J.B.J., 235
Freehand trend line, 13
Fundamental research, 2
Furtherest neighbour, 151

G

Gajar, 88
Gantt, Henry L., 217
Gathering knowledge for the sake of knowledge, 2
General Theory of Employment Interest and Money, 255
General theory, 186, 256
Geometrical analysis, 104
German anthropologist, 136
German, 181
Glass, G.V., 260, 261, 269
GNP, 199
Graduate student, 262
Grand centroid, 128
Greek letters, 49
Group centroid, 128
Group covariance natrices, 135
Grower, J., 144
Guilford, 264

H

Harvard Graduate School of Business Administration, 161
Hastini, 136
Hedges, 269
Herman, H.H., 116
Heterosedaticity, 50
Hierarchical agglomerative, 149
Hierarchical clustering method, 216
High speed computers, 138
Hindu, 69
Historians, 10
Homosedasticity, 50
Hoteling, H., 95
Hypothesis, 137
Hypothetical data, 132
Hypothetical variables, 95, 100

I

Idiosyncratic sample variations, 135
Image analysis, 109
Independent variable, 11, 38, 48, 49, 58, 63, 64, 65, 66, 67, 80, 82, 88, 103, 273
Inorganic chemistry, 138
Intercepts, 38
Interpretation of group differences, 123
Invisible hand, 175, 181
Iterative partitioning, 149

J

Jackson, J.B., 258, 260
Jevons, Stanlay W., 235

K

Kama Sutra, 136
Kelly, James, 217
Keynes, J.M., 255
Knowledge, 1
Kruskal, W.H., 92

L

Laboratory research, 2
Labour, 255
Laissez faire, 181
Law of value, 182
Least squares method, 109
Light, 273
Light, R.J., 259
Lincoln, 207
Linkage dendrograms, 152, 153
Literally analysis, 6
Logic restraints, 221
Love-life of human beings, 136

M

Mahalnobis, 134
Main group, 156, 157, 158
Management, 162
Man-machine system, 164
Marginal probabilities, 89
Marketing, 7
Marshall, Alfred, 6
Marx, Karl, 7, 181, 182, 252, 253
Mathematical analysis, 16
Mauchly, J., 217
Maximum likelihood analysis, 117
Maximum likelihood method, 109
Mekean, Ronald N., 162
Mental horizon, 136
Meta analysis, 8, 258-283
  abstracting results, 261-262
  adjusting effect size estimates, 268-269
  assessing research results, 263
  combining studies with focussed tests, 278-280
  comparison with diffused tests, 275-276
  counting method, 274
  critics of the analysis, 280-283

- meta analysis method mixes apples and oranges, 280-281
- meta analysis use data for poor studies, 281-282
- selection bias in recorded research, 282-283

defining study results, 266-268

development of meta analysis, 259-260

meta analysis of studies, 274-275
- combination of effect sizes, 272-273
- combining studies, 271
- effect size estimation, 271-272

quality of research results, 263-266

reliability of retrieval, 262

– – the literature, 261

retrieving research results, 260-261

strengths of meta analysis, 283

types of analysis, 259

voting method of integration, 273

when final results are not available, 269-270

Miller, 207, 261

MLR, 58

Mojena, R., 160

Monotonic, 154

Mostteller, F.M., 260

MQM, 94

Mriga (deer), 136

Multicollinearity, 66, 67, 68

Multiple correlation, 96

Multiple regression, 65, 69, 124

Multiple regression coefficient, 62, 64

Multiple regression model, 58

Multistage sampling, 5

Multivariate normal distribution, 117

Muslim, 69

Mutually independent, 250

N

National income, 29

Network analysis, 8, 188-216
- actor multiplicity, 211
- adjacent matrix, 194-195
- centrality index, 211-212
- characteristics of a graph, 189-190
- clique detection, 213
- directed graphs, 197
- distance measure, 214-216
- ego-network density, 211
- incidence matrix of a diagraph, 197-198
- incidence matrix, 193-194
- indices for networks, 204-206
- isomorphic graphs, 192
- levels of analysis, 202-203
  - complete network, 203
  - dyads, 202
  - ego-centric, 202
  - triads, 203
- matrix representation of graphs, 193
- network cohesion, 209
- – density, 208
- – multiplicity, 209-211
- path distance, 206-208
- paths, 192
- power of adjacency matrix, 195-196
- prestige index, 212-213
- principles of graph theory, 188-189
- relational content, 202
- research design, 199
- sampling, 200-202

  sociogram, 198
  some concepts of graph theory, 191
  structural equivalence, 213-214
  structure of networks, 203-204
    social cohesion, 203
    structural equivalence, 203-204
  sub-graphs, 192
  types of analysis, 199
  – – graphs, 190-191
  walks, 192
  X^3 matrix, 196
Network technique, 188
Nominal level, 68
Nominal variables, 71, 72, 76, 79, 80
Non-interval variables, 68
Null graph, 191
Null hypothesis, 52, 53, 54, 55, 56, 63, 64, 65, 74, 76
Number of time intervals, 19

O

Oblique factor model, 104
Oblique model, 107
  analysis, 109
Observed error, 88, 89, 90
One tailed test, 52
Operation of the central limit theorem, 52
OR, 121
Ordinal data, 85
Ordinal level, 68
Orthogonal model, 106, 108, 109
Orthogonal, 126
Oscillatory series, 19
OT, 121

P

Padmini, 136
Path distance matrix, 206
Pattern matrix, 107
Pattern of clustering by multivariate mixture analysis, 137
Pearson, Karl, 95, 138, 260, 266
Pearson's Product Moment Correlation, 82
Periodic series, 19
PERT, 217, 218, 224, 228, 229
Phi square, 82
Physical algebra, 238
Physical anthropology and biology, 122
Physical structure, 136
Physicians, 122
Pilot, 165
Piogu, 256
Planning, 162, 218
Planning Commission of India, 218
Polaris Missile Project, 218
Polithetic, 154
Political, 164
Political economy, 6, 7
Political scientists, 85
Pont, Du, 217
Population, 29, 40
Population correlation matrix, 115
Population regression coefficients, 43
Population Standard Deviation, 267
Price of commodities, 7
Price, 238
Principal Component Analysis, 95
Principal component method, 109
Principles of Economics, 6
Principles of Love, 136
Principles of numerical taxonomy, 136-137
Production of cement in India, 11
Production, 7
Programme Evaluation Research Task, 218

Psychological and education testing of children, 122
Psychologists, 10
Pumpkin, 88

Q

Qualitative data, 141
Qualitative variables, 70, 144
Quantitative analysis, 6
Quantitative variables, 70, 144
Quantity of money, 7
Quardant analysis, 94
Quota sampling, 5

R

RAND, 161
Rank of a matrix, 118
Rank theorem of matrices, 118
Raw variables, 95
Regression analysis, 10, 11, 37-69
  assumptions and inferences, 49-52
  calculation of regression coefficients, 43-46
  coefficient of determination, 46-47
  computation of the coefficients, 60-63
  detection of multicollinearity, 66-67
  dummy variables, 68-69
  error term, 40-42
  estimate of the coefficients, 59-60
  F-test, 63-65
  how to avoid multicollinearity, 67-68
  intercept, 38
  interpretation of R2, 47-49
  linear regression function, 58-59
  multicollinearity, 65-66
  multiple linear regression, 58
  one tailed test, 55
  prediction error of Y, 57-58
  prediction, 46
  principle of least squares, 42
  quick test, 55-56
  significance tests, 63
  slope, 38-40
  test of significance, 52-55
  when statistically not significant, 56-57
    inadequate sample size, 56
    restricted variance in X, 57
    specification error, 57
    type II error, 56-57
Regression coefficient, 38, 40, 42, 44, 79, 103
Regression curve, 42
Regression function, 42
Regression line, 46
Regression model, 16, 41, 46, 47, 55, 64
Regulation, 175
Research and development, 2
Research design, 281
Research findings, 281
Research methodology, 3
Residual matrix, 114
Rocky slope, 116
Rohlf, F., 137
Rosenthal, 262
Rosenthal, Robert, 260
Rotation of factor loadings, 120
RSD, 47
Rules of transformation, 172

S

Sample analysis, 49
Sample regression coefficients, 43
Sample, 5

Sankhini, 136
Sasa (hare), 136
Scheduling, 218
Schumpeter, J.A., 6-7
Schumpeter, J.M., 255
Science and technology, 161
Science of medicine, 138
Scientific, 1
Seasonal variation, 19, 22, 24
Self-regulated system, 164
Serial correlation coefficient, 36
Serial correlation, 35
Simple regression model, 59
Simplest parabola, 24
Single linkage method, 149
Slavery, 182
Smith, Adam, 6, 181, 261, 273
Smith, P.V., 259
Social behaviour, 136
Social circles, 203
Social scientists, 46
Sociologists, 10, 85
Sokal, R., 137
Spearman, Charles, 95
Spearman-Brown, 263, 266
Splinter group, 156, 157
Square symmetric matrix, 126
Standard deviation units, 128
Standard deviation, 53, 101
Standard error, 53, 56, 57, 66
Standard regression coefficients, 105
Standard variables, 101
Statistical analysis, 6, 16
Statistical correlation, 106
Stouffer method, 271
Stouffer, Samuel, 260
Strategic missile systems, 162
Stratified sampling, 5
Strength, 136
Structure Matrix, 106
Student, 199, 276
Sun, 164
Syam, 236
Systems analysis, 8, 161-187
  classification of systems, 164-165
  elements of a system, 170-171
  first application, 161-162
  function of a system, 169, 171-172
  input-output states, 170
  interconnections, 171
  Keynesian system, 186-187
  main characteristics, 163-164
  Marxian analysis of systems, 181-186
  Mason formula, 177-181
  regulation and control, 175-177
  response of a system, 169-170
  rule of multiplication, 172-173
  – – summation, 173-175
  – – operations, 172-175
  structure of a system, 168
  systems analysis in economics, 181
  – methodology, 165-168
    feedback, 166-167
    inputs, 165
    outputs, 165-166
    process, 166
    regulation and control, 167
    restrictions, 167
    systems theory, 167-168
  use in social sciences, 162

T

T matrix, 127, 128
Tatsuoka, M.M., 122
Taxonomy, 136

TE, 224, 225
Teacher, 199, 276
Technology, 172
Terminal vertex, 192
Theory of Political Economy, 235
Thompson, G.S., 116
Thorndike, 159
Tiedeman, D.V., 122
Time series analysis, 10, 36
  computation of serial correlation, 35
  cross-sectional data, 10
  differences of the series, 27-29
  elimination of trend and cycles, 25
  exponential trend line, 29
  finding of linear trend, 11-12
  free-hand method, 12
  measuring of seasonal variations, 21-23
  method of moving average, 15-18
  – – semi-average, 14
  modified exponential trend, 29-33
  other forms of trend lines, 24-27
  procedure of finding the trend line, 13-14
  quantitative deviations, 19
    oscillatory series, 19-20
    seasonal variations, 19
  ratio to moving average method, 23-24
  removal of I, 24
  seasonal and cyclical movement, 18
  – index, 26
  semi-average method, 33-34
  serial correlation, 34-36
  test of significance, 36
  time series data, 10-11
  – – regression model, 16
TL, 224, 225
Tools, 7
Total dispersion matrix, 127
Traite de Economie Pure, 235
Trend line, 22, 31
Trichotomised data, 72
TSD, 47
Tustin, 181
Two dimensional table, 72
Two-tailed test, 52, 55
Typology, 136
Tyron, R., 136

U

U.S. dollars, 237
UNIVAC Applications Research Centre, 217
US Air Force, 161
US Navy, 218
USA, 161, 217

V

Variance, 101

W

W matrix, 128
Walker, Morgan, 217
Ward, J.H., 153
Wealth of nations, 6
White, 200
Wiener, Norbert, 175, 181
Wolfe, J.M., 137
World War II, 161, 217

Y

Yin, R.K., 281

Z

Zubin, J., 136

Dr. Raghunath Padhy Sarma, Professor of Economics, Berhampur University (Rtd.) was born on 20.05.1936. He has done his M.A. (Eco.) in the year 1959 and Ph.D. in the year 1981. He joined as lecturer in Economics at Khallikote College, Berhampur in the year 1964 and retired as Professor from Berhampur University in the year 1996. He has guided eight Ph.D. scholars and at present four are under his guidance for Ph.D. work. He has published more than eighty articles at different journals of State and National level. He has also published five books. He is a good poet, painter, writer and also drama director. How he is engaged in various social activities. He is also able to speak and writes in Hindi, Telegu, Bengali, Assames, Sanskrit besides Oriya and English.

Dr. Rabi N. Misra completed his M.Com. in 1971, LL.B. in 1972, M. Phil in 1983. He got his Ph.D. degree in the year 1992 from Berhampur University. Dr. Misra has a large number of published works on UGC sponsored projects. He has published 14 books and four books are under print. He has published 60 articles in different journals.

Dr. Misra produced Five Ph.D. scholars and two scholars have submitted their Ph.D. thesis for award. At present he is guiding three scholars for Ph.D. work.

Dr. Misra is also good Oriya poet. He has published six Oriya poem books and two poem books are under print. He is associated with a number of voluntary organisations of the district and state and engaged in social activities. Dr. Misra is now working as professor in MBA in SMIT under Beju Patnaick University of Technology, Orissa.

ISBN 81-8356-110-1